RESEARCH

Christiane Behnert

Popularität und Relevanz in der Suche

Ein Experiment zur Erforschung von Relevanzkriterien in akademischen Suchsystemen

OPEN ACCESS

Springer Vieweg

Popularität und Relevanz in der Suche

Christiane Behnert

Popularität und Relevanz in der Suche

Ein Experiment zur Erforschung von Relevanzkriterien in akademischen Suchsystemen

Springer Vieweg

Christiane Behnert
Universität Hildesheim
Hildesheim, Deutschland

Die vorliegende Arbeit wurde vom Fachbereich 3 (Sprach- und Informationswissenschaften) der Universität Hildesheim als Dissertation mit dem Titel „Nutzerkriterien bei der Relevanzbewertung in akademischen Suchsystemen: Ein experimenteller Ansatz zur Erforschung von Relevanzkriterien am Beispiel von Popularitätsdaten als Bestandteil der Suchergebnispräsentation“ zur Erlangung des akademischen Grades einer Doktorin der Philosophie (Dr. phil.) angenommen.

Erster Gutachter: Prof. Dr. Joachim Griesbaum, Universität Hildesheim
Zweite Gutachterin: Prof. Dr. Christa Womser-Hacker, Universität Hildesheim
Dritte Gutachterin: Prof. Dr. Ulrike Spree, Hochschule für Angewandte Wissenschaften Hamburg

Tag der mündlichen Prüfung: 13. Dezember 2021

Hil 2

ISBN 978-3-658-37511-9 ISBN 978-3-658-37512-6 (eBook)
https://doi.org/10.1007/978-3-658-37512-6

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbiblio-grafie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

Planung/Lektorat: Stefanie Eggert
Springer Vieweg ist ein Imprint der eingetragenen Gesellschaft Springer Fachmedien Wiesbaden GmbH und ist ein Teil von Springer Nature.
Die Anschrift der Gesellschaft ist: Abraham-Lincoln-Str. 46, 65189 Wiesbaden, Germany

Vorwort

Nach Beendigung meiner Master-Arbeit im Studiengang Informationswissenschaften der Fachhochschule Potsdam konnte ich mir zunächst nicht vorstellen, eine Doktorarbeit zu schreiben und mich nicht nur sechs Monate, sondern mehrere Jahre intensiv mit einem speziellen Thema zu befassen. Dass ich die Motivation und das nötige Selbstvertrauen entwickelte, ein Promotionsvorhaben umsetzen zu wollen, verdanke ich dem glücklichen Umstand, im Rahmen des Forschungsprojekts *LibRank*[1] an der HAW Hamburg als Wissenschaftliche Mitarbeiterin einerseits mit dem Thema Relevanz „angesteckt" worden zu sein und andererseits das wissenschaftliche Forschen und Schreiben in der inspirierenden Umgebung der Forschungsgruppe Search Studies schätzen gelernt zu haben. Mein besonderer Dank gilt daher den Kolleginnen und Kollegen der Forschungsgruppe Search Studies am Department Information der HAW Hamburg während meiner Zeit als Wissenschaftliche Mitarbeiterin: Friederike Hanisch (geb. Kerkmann), Sebastian Sünkler, Sebastian Schultheiß, Prof. Dr. Dirk Lewandowski und Prof. Dr. Ulrike Spree. Als meine Betreuerin versorgte mich Ulrike Spree mit anregenden Diskussionen und ließ mir den nötigen Freiraum für meine persönliche Entwicklung auf dieser Promotionsreise. Dafür bin ich ihr sehr dankbar. Der HAW Hamburg danke ich für die dreijährige Promotionsförderung, ohne die ich das Promotionsvorhaben vermutlich nicht verfolgt hätte.

Mein Dank gilt ebenfalls Daniela Sygulla für ihre fleißige Mithilfe beim Erfassen der E-Mail-Adressen potenzieller Versuchspersonen. Ich bedanke mich vielmals bei Herrn Sebastian Ullrich von der punkt05 Statistikberatung für seine hilfreichen Hinweise zur Datenanalyse. Ebenfalls dankbar bin ich allen Mentorinnen und Mentoren für ihr konstruktives Feedback im Rahmen der internationalen

[1] https://gepris.dfg.de/gepris/projekt/246011126

Doktorandenkonsortien während der ISI 2017 in Berlin, des Annual Meetings der ASIST 2017 in Washington, D.C., der ISIC 2018 in Krakau und der CHIIR 2019 in Glasgow.

Schließlich gilt mein Dank Prof. Dr. Joachim Griesbaum und Prof. Dr. Christa Womser-Hacker dafür, dass sie mich im Rahmen des kooperativen Promotionsverfahrens als Doktorandin an dem Fachbereich Sprach- und Informationswissenschaften der Universität Hildesheim aufnahmen und mir ungeachtet der räumlichen Ferne viel Vertrauen entgegenbrachten. Herzlich danken möchte ich auch allen anderen Promovierenden in Hildesheim – stellvertretend sei hier Wiebke Thode genannt – für die vielen netten und ermutigenden Gespräche und die angenehme Atmosphäre bei den persönlichen Treffen, für die ich immer sehr gerne von Hamburg nach Hildesheim reiste.

Die Veröffentlichung meiner Dissertation wurde dank finanzieller Unterstützung der HAW Hamburg aus Mitteln des Open Access Publikationsfonds sowie aus Mitteln der Forschungsförderung der Universität Hildesheim ermöglicht. Ich freue mich über die Wertschätzung, die meine Doktorarbeit in dem so bedeutenden informationswissenschaftlichen Bereich der Relevanzforschung auf diese Weise erfährt und hoffe, dass alle an meiner Arbeit Interessierten durch die Lektüre Inspiration und hilfreiche Denkanstöße zur (experimentellen) Erforschung von Relevanzkriterien, Relevanzmerkmalen und Relevanzfaktoren erhalten.

Hamburg Christiane Behnert
im März 2022

Kurzfassung

Zur Bewertung der Relevanz von Suchergebnissen in Websuchmaschinen oder Bibliothekskatalogen verwenden informationssuchende Personen verschiedene Kriterien, anhand derer sie die Bewertung vornehmen. Moderne akademische Suchsysteme integrieren neben den traditionellen Metadaten zusätzliche Informationen über das Dokument in die Suchergebnisdarstellung, die für die Relevanzbewertung herangezogen werden können wie beispielsweise die Zitationszahl bei Google Scholar. Solche Popularitätsdaten können als Indikator für Qualität gesehen werden. Bisher gibt es keine Studien zu Relevanzkriterien, die Popularitätsdaten als Bestandteil von Such-ergebnissen im akademischen Kontext berücksichtigen. An dieser Stelle setzt die vorliegende Arbeit an, in dem sie die Kriterien bei der Relevanzbewertung von Suchergebnissen in akademischen Suchsystemen in den Fokus nimmt und den Einfluss von Popularitätsdaten auf die Bewertung untersucht. Zu diesem Zweck wurde zunächst ein Modell entwickelt, das die Einflüsse, die in dem Prozess der Relevanzbewertung von Suchergebnissen in akademischen Suchsystemen eine Rolle spielen, abbildet. Das Modell bietet erstmals eine systematische Übersicht über die Elemente eines Suchergebnisses als potenzielle Relevanzmerkmale, anhand derer informationssuchende Personen Relevanzkriterien ableiten, während dieser Prozess durch verschiedene Relevanzfaktoren beeinflusst wird. Damit stellt das Modell einen Lösungsvorschlag zur definitorischen und konzeptuellen Abgrenzung der in der Literatur oftmals unscharfen Begriffe Merkmale (*relevance clues/cues*), Kriterien (*relevance criteria*) und Faktoren (*relevance factors*) dar und dient zugleich als Hilfsmittel zur Operationalisierung von Relevanzkriterien.

Zur Untersuchung des Einflusses von Popularitätsdaten auf die subjektive Relevanzbewertung von Suchergebnissen in akademischen Suchsystemen wurde ein Experiment entwickelt, in dem die Anzahl von Downloads (UV 1), die

Zitationszahl des Werks (UV 2) und die Zitationszahl des Autors (UV 3) als unabhängige Variablen auf jeweils drei Stufen (gering, hoch und keine Angabe) in einem Within-Subjects-Design variiert wurden. Mithilfe eines Online-Fragebogens wurden von über 700 Teilnehmenden aus unterschiedlichen wissenschaftlichen Fachdisziplinen Relevanzbewertungen zu jeweils neun Surrogaten in drei Aufgaben erhoben, sodass pro Person 27 Bewertungen vorliegen. In die statistische Auswertung mit SPSS gingen die Daten von 627 Teilnehmenden ein. Die Ergebnisse der Mehrebenenanalyse liefern statistisch signifikante Haupt- und Interaktionseffekte, jedoch kein eindeutiges Bild über die Richtung der Effekte. Stattdessen zeigen sie ein komplexes Ergebnismuster, das die Komplexität des Relevanzkonzepts widerspiegelt.

Schlagwörter: Relevanz · Relevanzkriterien · Relevanzfaktoren · Relevanzbewertung · Surrogate · akademische Suchsysteme

Abstract

In order to judge the relevance of search results in web search engines or library catalogues, users apply various criteria to judge the relevance of their search results. Modern academic search systems integrate additional information about the document into the search result presentation, in addition to the traditional metadata, which can be used for relevance judgments, such as the citation count in Google Scholar. Such popularity data can be seen as an indicator for quality. So far, there are no studies on relevance criteria that consider popularity data as part of the search results in an academic context. This is where the present work sets in by focusing on the criteria for relevance judgments of search results in academic search systems and by investigating the influence of popularity data on judgments. For this purpose, a model was developed that shows the process of relevance judgment of search results in academic search systems and the influencing parameters that are part of this process. The model is the first to provide a systematic overview of the elements of a search result as potential relevance clues from which relevance criteria are derived, while various relevance factors influence this process. Thus, the model represents a proposed solution for the definitional and conceptual delimitation of the terms *relevance clues/cues*, *relevance criteria*, and *relevance factors*, and it serves at the same time as an aid for the operationalization of relevance criteria.

To investigate the influence of popularity data on the relevance judgment of search results in academic search systems, an experiment was developed in which the number of downloads (UV 1), the citation number of the work (UV 2), and the citation number of the author (UV 3) were varied as independent variables on three levels (low, high and not stated) in a within-subjects design. By using an online questionnaire, more than 700 participants from different scientific disciplines were asked to rate nine surrogates' relevance, each with the manipulated

popularity data in three different tasks, providing 27 judgments by every participant. The statistical analysis with SPSS included data from 627 participants. The multi-level analysis results provide statistically significant main and interactive effects, but do not give a clear picture of the direction of the effects. Instead, they show an intricate result pattern that reflects the complexity of the relevance concept.

Keywords: relevance · relevance criteria · relevance factors · relevance judgment · predictive judgments · search results · academic search systems

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Einleitung

1

Heutzutage ist das Nutzerverhalten bei der Informationssuche sehr stark durch die Websuche geprägt: Menschen berücksichtigen in der Regel nur die erste Suchergebnisseite und legen ihr Hauptaugenmerk auf die ersten drei Treffer, d. h. sie vertrauen dem Rankingalgorithmus der Websuchmaschine, die für sie relevanten Ergebnisse auf den obersten Positionen anzuzeigen (Asher et al., 2013; C. Barry & Lardner, 2011; Jansen & Spink, 2006; Nicholas et al., 2008; Pan et al., 2007; Schultheiß et al., 2018). Aus diesen Studien lässt sich schließen, dass das Relevanzranking die menschliche Relevanzbewertung maßgeblich beeinflusst. Doch nicht nur das Ranking der Suchergebnisse nimmt einen sehr großen Stellenwert bei der Informationssuche und der Auswahl der Suchergebnisse ein; auch deren Darstellung liefert Hinweise darüber, wie diese von informationssuchenden Personen bewertet werden. So zeigten beispielsweise Kammerer & Gerjets (2014), dass die Beurteilung der Glaubwürdigkeit von Informationsquellen durch die Darstellung der Suchergebnisse als Raster im Vergleich zur Präsentation als Liste positiv beeinflusst wird.

Zu verstehen, wie Menschen Informationen nach deren Relevanz bewerten, ist für die informationswissenschaftliche Forschung von großer Bedeutung. Einerseits stellt die Relevanzbewertung einen Teilprozess des gesamten Informationssuchprozesses dar, d. h., um den Informationssuchprozess in Gänze verstehen zu können, muss auch dieser Teilprozess erforscht werden. Zum anderen werden Relevanzbewertungen in der (Interactive-)Information-Retrieval (IR)-Forschung erhoben, um die Effektivität von Suchsystemen anhand von relevanzbasierten Kennzahlen zu messen.

Gegenstand der vorliegenden Arbeit ist der Prozess der Relevanzbewertung von Suchergebnissen in IR-Systemen. Dieser Prozess ist durch diverse Einflüsse gekennzeichnet, wodurch Relevanzbewertungen als Produkt des Bewertungsprozesses oft inkonsistent sind. So zeigten Buckley & Voorhees (2005), dass

C. Behnert, *Popularität und Relevanz in der Suche*,
https://doi.org/10.1007/978-3-658-37512-6_1

Bewertungen in Information-Retrieval-Studien zwischen Jurorinnen und Juroren – aber auch durch dieselbe Person zu unterschiedlichen Zeitpunkten – nicht zwangsläufig dieselben sind. Erklären lässt sich dies mit dem dynamischen, multidimensionalen und subjektiven Konzept von Relevanz, das in der Informationswissenschaft einerseits in der Informationsverhaltensforschung (*information behaviour research*), genauer im Bereich des Informationssuchverhaltens (*information searching behaviour*), andererseits im Bereich des (Interactive) Information Retrieval, (I)IR, theoretisch verortet ist. Vor diesem Hintergrund lässt sich die vorliegende Arbeit sowohl der Informationsverhaltensforschung als auch der (I)IR-Forschung zuordnen. Sie greift ein Kernkonzept der Informationswissenschaft auf, welches mit der Problemdarstellung im nachfolgenden Abschnitt 1.1 näher beleuchtet wird. Das Ziel dieser Arbeit, die bearbeiteten Forschungsfragen sowie deren Neuheitswert werden anschließend in Abschnitt 1.2 vorgestellt. Der Aufbau der Arbeit orientiert sich an dem gesamtmethodischen Vorgehen und wird in Abschnitt 1.3 dargelegt.

1.1 Problemdarstellung

Relevanz gilt bis heute als Kernkonzept der Informationswissenschaft (Greisdorf, 2000; Hjørland, 2000; Saracevic, 2016a, 2016b, 2015; White, 2009). Die Informationswissenschaft beschäftigt sich mit dem Relevanzkonzept, da ihm im Kontext der Suche und des Wiederauffindens von Informationen[1] eine zentrale Bedeutung zukommt: Das menschliche Verhalten während der Informationssuche ist motiviert durch den Wunsch, *relevante* Informationen zu finden; das Ziel von Information-Retrieval-Systemen besteht darin, *relevante* Suchergebnisse zu produzieren (Saracevic, 2015).

Unser Verständnis von Relevanz beruht auf Erkenntnissen aus jahrzehntelanger informationswissenschaftlicher, insbesondere Information-Retrieval-Forschung, die im Wesentlichen 1958 im Rahmen der International Conference on Scientific

[1] In dieser Arbeit wird der Informationsbegriff als „information-as-thing“ nach Buckland (1991) verstanden, da es grundsätzlich um Informationen als Objekte, d. h. in Verbindung mit einem Dokument als Träger von Informationen, geht, die von IR-Systemen gespeichert und (wieder)aufgefunden werden können. Diese Informationsobjekte werden als Suchergebnisse in Form von Dokumentsurrogaten der mit dem IR-System interagierenden, informationssuchenden Person von dem System angezeigt. Für eine angenehmere Lesbarkeit werden im Verlauf der Arbeit die Begriffe Dokument, Informationsobjekt, Informationen – wenn nicht anders beschrieben – synonym verwendet.

Information (ICSI) ihren Anfang nahm, als das Relevanzkonzept in den Arbeiten von Vickery (1959a, 1959b) zum ersten Mal im IR-Kontext diskutiert wurde (Mizzaro, 1997). Seit langem werden zwei grundsätzliche Perspektiven auf Relevanz eingenommen (Borlund, 2003b; Saracevic, 2007a): Relevanz aus Sicht des Systems (*system's view*), auch objektive oder algorithmische, logische Relevanz[2], und Relevanz aus Sicht des Nutzers (*user's view*), auch subjektive Relevanz. Allerdings wird die Dualität der beiden Sichtweisen seit längerem als kurzsichtig kritisiert (Mizzaro, 1997); Hjørland (2010) verlangt eine neue Interpretation der beiden Sichtweisen und Saracevic (2007a) plädiert für eine engere Verzahnung beider. Dass es schon lange nicht mehr ausschließlich die systembasierte Perspektive gibt, zeigt unter anderem die Etablierung des Interactive Information Retrieval (IIR) als Forschungsfeld, das die Interaktion der suchenden Person mit dem System in den Fokus nimmt (Belkin, 2015; Borlund, 2013; Cool & Belkin, 2011; Ruthven, 2008).

Die traditionelle Information Retrieval (IR)-Forschung betrachtet Relevanz von der Systemseite her und zielt auf die Ermittlung der Relevanz zwischen einer Suchanfrage und einem Dokument ab, d. h. dieser Betrachtung liegt eine anfrageorientierte Sichtweise zugrunde. Dabei wird ein Dokument anhand des Grads der thematischen Übereinstimmung mit der Suchanfrage als (thematisch) relevant oder irrelevant angesehen (Baeza-Yates & Ribeiro-Neto, 2011; Borlund, 2003b; Dervin & Nilan, 1986). Was aus Systemsicht als thematisch relevant bewertet wird, ist allerdings vielmehr mit *Aboutness* zu beschreiben und nicht gleichzusetzen mit thematischer Relevanz aus Nutzersicht; Aboutness beschreibt den thematischen Gegenstand des Dokumenteninhalts und wird einem Dokument zugeschrieben unabhängig von einer Suchanfrage; somit ist Aboutness als statisch anzusehen, im Gegensatz zu Relevanz, die als hoch dynamisch gilt (Saracevic, 2012).

Gegen eine exklusiv anfrageorientierte Sichtweise argumentierte Bookstein (1979): „For one, it is the patron, not the request, that is being served, so a request-oriented definition has no practical content. Also, there is no way of

[2] Der Begriff objektive Relevanz führt auf die Definition von Relevanz im IR-Kontext von Cooper (1971) zurück. Cooper schlug vor, Relevanz als logische Relevanz (*logical relevance*) zu betrachten, wobei er zugleich die Grenzen betonte, auf die eine rein logische Relevanz trifft, da sie auf einen rein binären Kontext in Hinblick auf die Art der Suchanfrage und die Möglichkeiten des Systems, diese Suchanfrage zu beantworten, beschränkt ist; zudem erfordert sie eine formalisierte Anfragesprache. Aus heutiger Sicht geht systembasierte Relevanz über die rein logische Relevanz hinaus, denn sie beruht auf der Übereinstimmung von Suchtermen auch aus natürlichsprachlichen Suchanfragen mit Begriffen, die in den Metadaten bzw. dem Volltext eines Dokuments vorkommen.

asking the request what it thinks about the document“ (S. 270). Tatsächlich setzt die Bestimmung von thematischer Relevanz (*topical relevance, topicality*) zwischen einer Suchanfrage und einem Suchergebnis Wissen über Begriffe und Konzepte, die ein Thema repräsentieren, voraus (Hjørland & Christensen, 2002); sie erfordert daher einen komplexen kognitiven Prozess (Huang & Soergel, 2013), wodurch thematische Relevanz nicht der systemseitigen, sondern eher der nutzerseitigen Sicht zuzuordnen ist. Thematische Relevanz kann allerdings oft mithilfe des vom System erzeugten Ergebnisses abgeleitet werden (Saracevic, 1996).

Vertreterinnen und Vertreter der Nutzerseite hingegen betrachten Relevanz aus der Perspektive der suchenden Person unter Berücksichtigung ihres Kontexts, insbesondere ihres Informationsbedürfnisses in der jeweiligen Situation (Bookstein, 1979; Cosijn & Ingwersen, 2000; Kemp, 1974; Schamber et al., 1990; P. Wilson, 1973) und den kognitiven Veränderungen des Menschen, die mit Voranschreiten des Suchprozesses einhergehen (Belkin, 1980; Harter, 1992). Gegen eine reine Nutzerperspektive allerdings spricht, dass Nutzerinnen und Nutzer auch mit weniger als optimalen Suchergebnissen zufrieden sein können (siehe *satisficing,* Case & Given, 2016, S. 36; Mansourian & Ford, 2007; Prabha et al., 2007; Savolainen, 2016) und die Sicht der einzelnen Person zum Zeitpunkt der Suche nicht zwangsläufig die relevanteste ist (Hjørland, 2010, S. 223).

Der Relevanzbegriff steht für ein komplexes Konzept, das vom Menschen intuitiv verstanden wird (Saracevic, 1996), wohingegen das Ziel von IR-Systemen darin besteht, die Nutzersicht auf Relevanz abzubilden – „All the algorithms in all the systems in the world are trying to approximate, with various degrees of success, the human notion of relevance“ (Saracevic, 2012, S. 49). Zu diesem Zweck werden nutzungsbasierte Faktoren (Signale des menschlichen Suchverhaltens) in den Ranking-Algorithmus integriert, was sich seit der Einführung des PageRank-Verfahrens[3] durch die Erfinder von Google (Brin & Page, 1998) etabliert hat.

Für die Evaluierung der Retrieval-Effektivität von IR-Systemen werden menschliche Relevanzbewertungen im Rahmen von Retrieval-Studien erhoben. Anhand dieser Relevanzbewertungen werden auch heutzutage relevanzbasierte Kennzahlen (z. B. *Normalised Discounted Cumulative Gain,* NDCG, die die Treffersortierung berücksichtigt, oder *Graded Average Precision,* GAP) errechnet. Aus diesen Gründen braucht die Systemseite die Nutzerseite und profitiert von ihr. Abbildung 1.1 veranschaulicht die Arten von Relevanzbewertungen, nach

[3] Das Konzept hinter dem PageRank-Verfahren folgt dem Prinzip der „Weisheit der Vielen“ (Surowiecki, 2005): Je häufiger auf eine Webseite verlinkt wurde und je populärer im Sinne des PageRank-Algorithmus die verlinkenden Dokumente sind, desto wahrscheinlicher ist es, dass diese Webseite für andere Personen als relevant gesehen wird.

denen sie in Abhängigkeit der system- oder nutzerbasierten Perspektive unterschieden werden können. Aus Systemsicht sind Relevanzbewertungen formalisierte *relevance scores,* die der Rankingalgorithmus errechnet; aus Nutzersicht sind Relevanzbewertungen das Ergebnis eines kognitiven Bewertungsprozesses.[4] Dabei besteht die Annahme, dass die Bewertungen, die informationssuchende Personen in ihrem beruflichen oder privaten Kontext (Alltag) vornehmen, den im Forschungskontext erhobenen Relevanzbewertungen entsprechen. Erstgenannte können lediglich implizit anhand von Logdaten, wie beispielsweise Klickdaten und Verweildauer, abgeleitet werden (Agichtein et al., 2006; Joachims et al., 2005), während letztere mithilfe einer zwei- oder mehrstufigen Skala auch explizit erfassbar sind.

Dass sich Relevanzbewertungen je nach Perspektive unterschiedlich definieren lassen, spiegelt sich in der Forschungsliteratur wider[5]. Mizzaro (1997) definiert Relevanzbewertungen (*relevance judgments*) aus Sicht menschlicher Juroren („by a judge"), sodass von einer Definition im Forschungskontext ausgegangen werden kann: „A relevance judgment is an assignment of a value of relevance […] by a judge at a certain point of time" (S. 812).

Harter (1992) kritisiert, Relevanzbewertungen, die im Rahmen von Studien zur Evaluierung der Retrieval-Effektivität erhoben werden, als statische Grundlage zur Berechnung von Retrieval-Kennzahlen zu behandeln. Er begründet dies damit, dass dieser Ansatz der traditionellen, anfrageorientierten Systemsicht folgt und seiner Definition von psychologischer Relevanz als Beziehung zwischen dem psychologischen Zustand einer informationssuchenden Person zu einem bestimmten Zeitpunkt und dem Informationsobjekt widerspricht:

---

[4] Die vorliegende Arbeit betrachtet Relevanz und den Prozess der Relevanzbewertung ausschließlich aus menschlicher Perspektive. Die systemseitige Sicht auf Relevanz ist nicht Gegenstand dieser Arbeit; die Ermittlung von *relevance scores* anhand von Signalen über Nutzungsverhalten, die in Ranking-Algorithmen von IR-Systemen wie Websuchmaschinen zum Tragen kommen, findet keine Berücksichtigung.

[5] Eine wesentlich weiter gefasste Definition von Relevanzbewertungen gibt Hjørland, indem er Relevanzbewertungen bereits bei der Auswahl und Indexierung von Dokumenten für eine Datenbank verortet (Hjørland, 2001, S. 777). Da diese Art von Bewertung nicht im Rahmen eines individuellen Informationssuchprozesses stattfindet, gibt es jedoch kein subjektives Informationsbedürfnis, das es zu befriedigen gilt – allenfalls das Bedürfnis der Institution (z. B. Bibliothek) oder des Datenbankproduzenten bzw. -anbieters, Dokumente entsprechend bestimmter Vorgaben (z. B. Leitbild, Erwerbungsprofil, gesetzlicher Sammlungsauftrag) auszuwählen. Eine andere Definition überträgt die Relevanzbewertung sogar auf den gesamten Forschungsprozess und den Zeitpunkt, zu dem Forschende ihre Ergebnisse publizieren: „When authors cite other authors' work, they are making relevance judgments, linking ideas for future readers" (Schamber, 1994, S. 28).

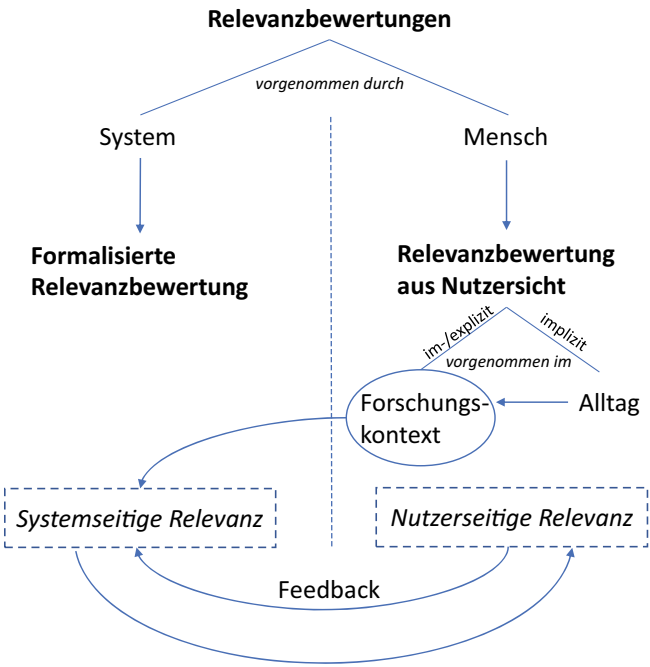

Abbildung 1.1 Unterscheidung von Relevanzbewertungen nach Relevanzperspektive

> A definition of relevance that relies on fixed-for-all-time, unchanging relevance judgments – such as those characterizing nearly all retrieval tests that have been conducted to now – must be seen as wrong. For relevance judgments are a function of one's mental state at the time a reference is read. They are not fixed; they are dynamic. Recording such judgments, treating them as permanent, unchanging relations between a document set and a question set, and then using them to compute such measures as recall and precision to evaluate retrieval effectiveness, is contrary to the meaning of psychological relevance. (S. 612)

Harters Argumentation folgend kann infrage gestellt werden, inwieweit explizite Relevanzbewertungen realen Relevanzbewertungen entsprechen können bzw. ob der Gegenstand der Bewertung im Forschungskontext überhaupt eine *Relevanz*beziehung sein kann. Andere Definitionen betonen den Zusammenhang zwischen Nutzerinnen und Nutzern („users", „the human user") und deren Informationsbedürfnissen:

> From our perspective, *relevance judgments* are users' evaluations of information (from internal or external sources) in relation to their information need situations at particular points in time. (Schamber u. a., 1990, S. 771; Kursivdruck im Original)
> Relevance judgments can be seen as a subjective process where the human user decides on the relevance of retrieved documents in relation to the work task that he/she has to fulfill. (Cosijn, 2009, S. 4513)

Die Definition von Cosijn geht darüber hinaus auf den subjektiven Prozesscharakter ein, der im Kontext von Relevanzbewertungen durch menschliche Nutzerinnen und Nutzer zwangsläufig vorliegt. Dieser Prozess der Relevanzbewertung unterliegt verschiedenen Einflüssen, die mitunter dazu führen, dass Inkonsistenzen bei expliziten, im Forschungskontext erhobenen Relevanzbewertungen auftreten: So sind Relevanzbewertungen zwischen Jurorinnen und Juroren (Interrater-Reliabilität, *inter-rater reliability*) sowie bei wiederholten Bewertungen durch dieselbe Person (Intrarater-Reliabilität, *intra-rater reliability*) nicht immer gleich (Buckley & Voorhees, 2005, S. 68 ff.), was ein starkes Argument für das kontextabhängige und dynamische Relevanzverständnis ist (Saracevic, 2016b, S. 69 ff.).

Zusammengefasst stellen Relevanzbewertungen das Produkt eines menschlichen, kognitiven Beurteilungsprozesses dar. Dieser Prozess kann wiederum als sequenzielle Anwendung von Relevanzkriterien definiert werden: „A relevance judgement process is […] defined as the sequential use of relevance criteria as delimited by interactions" (Beresi et al., 2010, S. 199). Um den Prozess der Relevanzbewertung zu verstehen, ist es demnach unerlässlich, detaillierte Kenntnisse über Relevanzkriterien aus der Perspektive von Nutzerinnen und Nutzern zu besitzen.

Viele informationswissenschaftliche Studien untersuchten gezielt, anhand welcher Kriterien Menschen Relevanzbewertungen vornehmen. Als einige der einflussreichsten Untersuchungen gelten die Arbeiten von Carol Barry und Linda Schamber (Barry, 1994; Barry & Schamber, 1998; Schamber, 1991), die unter anderem zeigen, dass thematische Relevanz die Basis für die Relevanzbewertung darstellt und dass weitere Kriterien wie Qualität, Validität und Aktualität auf dieser Basis aufsetzen. An diese Erkenntnisse knüpfen weitere Studien an und untersuchen teilweise gezielt ausgewählte Kriterien insbesondere vor dem Hintergrund der zunehmenden Nutzung von Websuchmaschinen. Besonders hervorzuheben ist die Arbeit von Rieh (2002) zur Bewertung von Qualität und kognitiver Autorität bei der Websuche – zwei Konzepten, die bei der Suche nach wissenschaftlichen Informationen, d. h. im Kontext akademischer Informationssuche, einen besonderen Stellenwert einnehmen dürften (Rieh & Belkin, 1998).

Das Konzept der kognitiven Autorität besagt, dass eine Person nicht nur als Experte gilt, sondern auch eine kognitive Autorität ist, wenn ihre Aussagen von anderen als wahr akzeptiert werden und sie zugleich Denken und Handeln anderer Menschen beeinflusst (Wilson, 1983). Die Einschätzung der kognitiven Autorität von Autorinnen und Autoren basiert auf deren Ruf und Leistungen (Rieh, 2009). Kennzahlen für den Einfluss von Forschenden bzw. deren Publikationen sind u. a. als Gegenstand des informationswissenschaftlichen Teilgebiets der Bibliometrie bzw. der Szientometrie – und mit der Forderung nach sogenannten alternativen Metriken (*altmetrics*[6]) in der Wissenschaftsevaluierung – auch disziplinübergreifend bekannt.

Bei der Informationssuche in akademischen Suchsystemen kann das Wissen über den Einfluss der Autorin oder des Autors hilfreich sein für die Bewertung der Ergebnisse durch die informationssuchende Person. Moderne akademische Suchsysteme integrieren zusätzliche Daten in ihre Ergebnispräsentation: Die Ergebnisse von Google Scholar enthalten zum Beispiel die Anzahl der Zitationen, die ACM Digital Library zeigt zusätzlich die Anzahl der Downloads des jeweiligen Werks in der Suchergebnisliste an. Obwohl Autorität als beeinflussendes Element bei der Relevanzbewertung in akademischen Suchsystemen bereits in einer Studie zur Dokumentenauswahl (Wang, 1994) in Betracht gezogen wurde, gibt es zum Zeitpunkt der Erstellung dieser Arbeit keine Studien zu Relevanzkriterien, in denen den Teilnehmenden Suchergebnisse mit angereicherten Daten wie Popularitätsdaten zur Bewertung vorgelegt wurden. Daher stellt sich die Frage, inwieweit die Ergebnisse der früheren Studien zu Relevanzkriterien auf den heutigen Kontext moderner akademischer Suchsysteme übertragen werden können, wenn bestimmte Elemente eines Suchergebnisses keine Berücksichtigung in damaligen Untersuchungen fanden, weil sie zu dem Zeitpunkt noch nicht existierten.

Bisherige Studien zeigen neben den von informationssuchenden Personen angewendeten Kriterien (*relevance criteria*) die Existenz weiterer Einflüsse auf den gesamten Bewertungsprozess auf, welche aus Einflussfaktoren (*relevance factors*) und Merkmalen oder Hinweisen auf Relevanz (*relevance clues, relevance cues*) bestehen (siehe z. B. Literaturüberblicke bei Mizzaro, 1997; Saracevic, 2016b). Zwischen den Begriffen Kriterien, Faktoren und Merkmalen gibt es in der Literatur weder eine definitorische und konzeptuelle Abgrenzung noch eine allgemeingültige Definition. Dadurch bedingte terminologische Differenzen erschweren die Einordnung und Vergleichbarkeit der Ergebnisse aus verschiedenen empirischen Studien zu Relevanzkriterien (Bales & Wang, 2006; Saracevic,

[6] http://altmetrics.org/manifesto/ (letzter Zugriff: 06.03.2021).

2016b; Wang, 2010). So listen manche Studien beispielsweise Aktualität als Einflussfaktor auf (z. B. Schamber, 1994), während andere Aktualität als Kriterium ausweisen (z. B. Barry, 1994).

Inhaltlich vergleichbare Primärstudien zu derselben Fragestellung oder demselben Effekt sind von besonderer Bedeutung für die Durchführung von Metaanalysen, die die Studienergebnisse zusammenfassen und statistisch auswerten; Metaanalysen bzw. Forschungssynthesen sind sowohl für den wissenschaftlichen Erkenntnisfortschritt als auch für die Anwendbarkeit wissenschaftlicher Erkenntnisse in der Praxis notwendig (Döring & Bortz, 2016, S. 894 ff.). Des Weiteren ist für eine empirische, quantitative Untersuchung von Relevanzkriterien die Operationalisierung der Untersuchungsvariablen unabdingbar. Ohne eine Definition dessen, was unter einem Relevanzkriterium zu verstehen ist, lässt sich dieses Konstrukt als Untersuchungsgegenstand nicht angemessen operationalisieren, d. h. die Definition muss der Operationalisierung vorausgehen, bevor sich an diese die Datenerhebung anschließt (Döring & Bortz, 2016, S. 222 ff.).

Bei den bisherigen empirischen Studien zu Relevanzkriterien handelt es sich zumeist um explorative Studien, die offene Forschungsfragen anhand eines qualitativen Ansatzes bearbeiteten (z. B. Fitzgerald & Galloway, 2001; Schamber, 1991; Wang, 1994; I. Xie et al., 2010), wesentlich seltener um hypothesenprüfende, quantitative Studien, denen ein experimentelles Design zugrunde liegt (z. B. Hamid et al., 2016; Regazzi, 1988; Xu & Chen, 2006). Ein Experiment ist jedoch die einzige Möglichkeit, um kausale Schlussfolgerungen zwischen einer vermuteten Ursache und einer beobachteten Wirkung ableiten zu können (Döring & Bortz, 2016, S. 192). Experimentelle Studien gelten daher als Goldstandard in der sozialwissenschaftlichen Forschung (Döring & Bortz, 2016, S. 102) und kommen insbesondere zum Erforschen von menschlichem Erleben und Verhalten in der Psychologie zum Einsatz.

Die informationswissenschaftlichen Literatur zu Relevanz und Relevanzkriterien liefert ebenfalls Hinweise auf den Verhaltensaspekt: So ist in einigen Publikationen die Rede von der Erforschung des Relevanzverhaltens (*relevance behavior/behaviour*) (z. B. Balatsoukas et al., 2010; Borlund, 2003b; Cook, 1971; Ruthven, 2014; Saracevic, 2007b, 1996; Scholer et al., 2013; Wang, 2011), auch die Bezeichnung Relevanzbewertungsverhalten (*relevance judgment behaviour*) (Balatsoukas & Demian, 2009; Balatsoukas & Ruthven, 2012) findet Verwendung. Saracevic erläutert den Begriff Relevanzverhalten im Kontext der Informationsverhaltensforschung:

> Strictly speaking, relevance does not behave. People behave. A number of studies examined a variety of factors that play a role in how humans determine relevance of information or information objects. Relevance behavior studies are closely related to information seeking studies and to the broad area of human information behavior studies. […] Many studies on various aspects of human information behavior are related to relevance behavior... (2007b, S. 2127)

Der subjektive Prozess der Relevanzbewertung als Teil des menschlichen Informationssuchprozesses ist als Gegenstand der *Information Searching Behavior*-Forschung und zugleich als Relevanzverhalten anzusehen. Somit erscheint es naheliegend, methodisch dem Ansatz der Psychologie zur Erforschung menschlichen Verhaltens zu folgen und den Bewertungsprozess als Anwendung verschiedener Relevanzkriterien anhand eines experimentellen Designs zu untersuchen.

1.2 Ziel und Relevanz der Arbeit

Das Ziel der vorliegenden Arbeit ist es, Kenntnisse über die Kriterien, anhand derer informationssuchende Personen die Relevanz von Suchergebnissen in akademischen Suchsystemen bewerten, zu erlangen. Dieses Ziel stellt die übergeordnete Forschungsfrage dar, welche die hier beschriebene Forschung anleitet.

Das in dieser Arbeit behandelte Forschungsproblem (vgl. Abschnitt 1.1) lässt sich wie folgt zusammenfassen: In den bisherigen – explorativen wie experimentellen – Studien zu Relevanzkriterien, die Relevanzbewertungen auf der Basis von Suchergebnissen erhoben, enthielten die von den Teilnehmenden zu bewertenden Suchergebnisse keine Popularitätsdaten, wie sie heutzutage in Form von Informationen über Download- oder Zitierhäufigkeit in modernen akademischen Suchsystemen als integraler Bestandteil der Ergebnispräsentation zu finden sind. Daher ist ungewiss, in welcher Weise solche Popularitätsdaten die Relevanzbewertung beeinflussen und welche Bedeutung der Popularität als Relevanzkriterium zugesprochen werden kann.

An dieser Stelle setzt die vorliegende Arbeit an, deren Hauptziel darin besteht, den Einfluss von Popularitätsdaten auf die Relevanzbewertung von Suchergebnissen in akademischen Suchsystemen empirisch zu ermitteln. Zu diesem Zweck wird ein Experiment entwickelt, denn nur mit Experimenten ist es möglich, kausale Schlussfolgerungen über Zusammenhänge zwischen Ursache und Wirkung abzuleiten. Die Frage nach dem Zusammenhang zwischen Popularitätsdaten als Bestandteil von Suchergebnissen (Surrogaten) in akademischen Suchsystemen (Ursache) und der Relevanzbewertung ebendieser Suchergebnisse (Wirkung) führt zu insgesamt drei Forschungsfragen, die mit dieser Arbeit beantwortet werden

sollen. Aufgrund fehlender vergleichbarer Studien zur experimentellen Erforschung von Relevanzkriterien bei der Bewertung von Suchergebnissen erfolgt zunächst die Auseinandersetzung mit den methodischen Anforderungen an die empirische Studie:

F1 Wie können Nutzerkriterien bei der Relevanzbewertung anhand eines experimentellen Untersuchungsdesigns erforscht werden?

Für die Entwicklung eines experimentellen Designs zur Erforschung von Relevanzkriterien sind zwei wesentliche Voraussetzungen zu erfüllen, die mit den folgenden Unterforschungsfragen bearbeitet werden:

F1a Wie lassen sich Merkmale, Kriterien und Faktoren als Einflüsse im Prozess der Relevanzbewertung für die Entwicklung eines experimentellen Untersuchungsdesigns definitorisch und konzeptuell voneinander abgrenzen?

F1b Wie können Kriterien bei der Relevanzbewertung von Suchergebnissen für eine experimentelle Studie operationalisiert werden?

Bisherigen Studien, in denen die nutzerseitigen Kriterien zur Relevanzbewertung erforscht wurden, mangelt es an einer eindeutigen und allgemeingültigen Definition dessen, was als Relevanzkriterium zu verstehen ist. Diese fehlende, kontextunabhängige Definition geht mit einer Vermischung von weiteren Begriffen als die Relevanzbewertung beeinflussenden Größen einher. So werden in der Literatur ebenfalls die Begriffe Relevanzfaktoren und Relevanzmerkmale gebraucht, deren Elemente eine klare Abgrenzung vermissen lassen. Diese Abgrenzung wird anhand eines Modells zur subjektiven Relevanzbewertung grafisch dargestellt. Dabei liegt der Fokus auf Surrogaten, die als Ergebnisse von informationsorientierten Anfragen, also thematischen Suchen, von einem textbasierten Information Retrieval-System produziert werden und die in der Regel mehrere Dokumente zur Befriedigung des Informationsbedürfnisses benötigen, im Gegensatz zu navigationsorientierten Suchanfragen, die gezielt ein (bekanntes) Informationsobjekt[7] verlangen und damit eine Relevanzbewertung im Sinne einer subjektiven Beurteilung überflüssig machen[8]. Den konkreten Kontext bilden dabei akademische Suchsysteme, die zusätzliche Daten wie Popularitätsdaten

[7] Die Einteilung von Suchanfragen im Web in informations-, navigations- und transformationsorientierte Anfragen geht zurück auf Broder (2002).

[8] Der Grund hierfür liegt in der Tatsache, dass Suchergebnisse navigationsorientierter Anfragen dichotom nach ihrer „Richtigkeit“ beurteilt werden können – entweder das Ergebnis ist

in die Suchergebnispräsentation integrieren. Dies wird durch das Modell explizit berücksichtigt, wodurch es eine zeitgemäße Sichtweise auf die Relevanzbewertung von Surrogaten in akademischen Suchsystemen bietet. Zudem stellt das Modell ein Hilfsmittel für die Operationalisierung von Relevanzkriterien dar, die für deren experimentelle Erforschung im Allgemeinen und im Rahmen der hier vorgestellten Studie notwendig ist. Mit der Beantwortung der Unterforschungsfragen F1a und F1b geht das Erreichen zweier Teilziele einher, die in einem für die Beantwortung der Forschungsfrage F1 entwickelten methodischen Framework münden, welches wiederum ein essenzielles Teilergebnis der hier beschriebenen Forschung darstellt.

Die inhaltlichen Erkenntnisse aus der experimentellen Studie zielen auf die Ermittlung des Einflusses von Popularitätsdaten auf die Relevanzbewertung ab:

F2 Welchen Einfluss haben Popularitätsdaten auf die Bewertung der Relevanz von Suchergebnissen in akademischen Suchsystemen?

Die Forschungsfrage F2 lässt sich in zwei Schritten beantworten. Zunächst ist zu prüfen, ob ein statistisch signifikanter Effekt vorhanden ist, bevor die Richtung des Effekts festgestellt werden kann. Ist der Effekt positiv, bedeutet dies, dass Popularitätsdaten zu einer höheren Relevanzbewertung führen; ein negativer Effekt liegt hingegen vor, wenn Popularitätsdaten eine geringere Relevanzbewertung bewirken. Forschungsfrage F3 nimmt die Effekte der untersuchten Popularitätsdaten einzeln in den Fokus:

F3 Welche Popularitätsdaten beeinflussen die Relevanzbewertung in welchem Maße?

Unterschiedliche Einflüsse bei den untersuchten Popularitätsdaten liefern Erkenntnisse über deren jeweiligen Stellenwert und können als Basis für Schlussfolgerungen über die Gewichtung von Popularität als Relevanzkriterium dienen. Die Beantwortung der Forschungsfrage F3 setzt voraus, dass ein statistisch signifikanter Einfluss von Popularitätsdaten nachgewiesen wird.

Mit der Bearbeitung dieser Forschungsfragen leistet die vorliegende Arbeit einen Beitrag zum besseren Verständnis des Prozesses der Relevanzbewertung. Da die Arbeit diesen als Teilprozess der Informationssuche betrachtet, unterstützt sie damit nicht nur die informationswissenschaftliche Relevanzforschung,

---

korrekt und es handelt sich um das gesuchte Dokument oder das Ergebnis ist nicht korrekt, weil es sich nicht um das gesuchte Dokument handelt.

sondern indirekt auch die Informationsverhaltensforschung und die Interactive Information Retrieval-Forschung. Zugleich möchte sie mit dem verfolgten experimentellen Ansatz zur Erforschung von Relevanzkriterien einen methodischen Impuls für weitere Studien setzen und bietet einerseits mit dem Modell zur Relevanzbewertung von Suchergebnissen in akademischen Suchsystemen, andererseits mit dem methodischen Framework zwei nachnutzbare Hilfsmittel für die Operationalisierung von Variablen und die Entwicklung experimenteller Untersuchungsdesigns.

Der Neuheitswert der vorliegenden Arbeit für die Informationswissenschaft liegt zum einen in dem methodischen Framework, anhand dessen erstmalig eine experimentelle Studie zur Untersuchung des Einflusses von Popularitätsdaten auf die Relevanzbewertung entwickelt wird. In diesem Zusammenhang ist die Durchführung der Studie als Online-Experiment zu betonen, denn diese ermöglicht unter anderem, eine im Vergleich zu bisherigen Studien zu Relevanzkriterien sehr große Stichprobengröße zu erreichen. Zum anderen besteht der Neuheitswert in der erstmaligen Untersuchung des Einflusses von Popularitätsdaten, wie der Anzahl von Zitationen einer Autorin oder eines Autors oder die Anzahl der Downloads zu einem Zeitschriftenartikel, auf die Relevanzbewertung von Suchergebnissen in akademischen Suchsystemen.

1.3 Vorgehensweise und Aufbau der Arbeit

Die vorliegende Arbeit verfolgt einen quantitativen, experimentellen Ansatz zur Erforschung von Relevanzkriterien. Dabei stützt sie sich zunächst auf bestehende Forschungsliteratur, um das Konzept von Relevanz und Relevanzkriterien inhaltlich zu durchdringen sowie das methodische Vorgehen bei der Untersuchung von Relevanzkriterien bisheriger Studien zu beleuchten. Für die Entwicklung eines experimentellen Designs zur Erforschung von Relevanzkriterien müssen Voraussetzungen in Hinblick auf die Operationalisierbarkeit des zu untersuchenden Gegenstands erfüllt sein, die auch dessen Definition beinhalten. Um diese Voraussetzungen zu erfüllen, wird mithilfe eines literaturorientierten Ansatzes ein Modell entwickelt, welches den subjektiven Prozess der Relevanzbewertung von Suchergebnissen in akademischen Suchsystemen darstellt. Anhand des Modells werden die Einflüsse, die in dem Bewertungsprozess eine Rolle spielen, veranschaulicht und deren Zusammenwirken beschrieben. In diesem Zusammenhang werden die Forschungsfragen F1a und F1b beantwortet, ohne die die Entwicklung eines Experiments zur Erforschung von subjektiven Relevanzkriterien und die Beantwortung der Forschungsfrage F1 nicht möglich sind.

Den Kern dieser Arbeit bildet die experimentelle Studie, mit der der Einfluss von Popularitätsdaten als integraler Bestandteil der Suchergebnispräsentation in akademischen Suchsystemen untersucht wird. Aufgrund fehlender vergleichbarer, experimenteller Studien kann bei der Entwicklung des Designs und insbesondere der Erstellung des Stimulusmaterials nicht auf bereits vorhandene methodische Frameworks zurückgegriffen werden. Aus diesem Grund liegt ein Schwerpunkt der Studie in der detaillierten Beschreibung des methodischen Vorgehens, für das jede Entscheidung in Hinblick auf das Design und die Datenerhebung sowie die Datenauswertung ausführlich begründet wird. Die statistische Analyse der erhobenen Daten erfolgt mithilfe der Software SPSS, mit der eine Mehrebenenanalyse (*multi-level analysis*) durchgeführt wird, welche die statistischen Signifikanzen von Haupteffekten und Interaktionseffekten aufdeckt. Aufgrund des komplexen Designs nimmt die Darstellung der statistischen Ergebnisse des Experiments entsprechend viel Raum ein, daher ist eine Vielzahl an Tabellen separat im Anhang im elektronischen Zusatzmaterial enthalten. Zum Zweck der Transparenz dieser Forschung und zur Ermöglichung von Replikationsstudien sind die im Rahmen des Online-Experiments erhobenen Forschungsdaten im Open Science Framework hinterlegt und unter der folgenden URL abrufbar: https://doi.org/10.17605/OSF.IO/NTWQD.

Der strukturelle Aufbau der Arbeit folgt dem Vorgehen zur Bearbeitung der Forschungsfragen. Zunächst wird in **Kapitel 2** der Stand der Forschung dargelegt, der sich in zwei Teile gliedert. Der erste Teil (Abschnitt 2.1) beleuchtet die Kriterien bei der Relevanzbewertung aus einer inhaltlichen Perspektive heraus, wobei die Kriterien in allgemeine Kriterien (Abschnitt 2.1.1), Kriterien im Kontext der Websuche (Abschnitt 2.1.2) und in Kriterien im Kontext akademischer Suchsysteme (Abschnitt 2.1.3) unterteilt werden. Schließlich wird der Fokus auf Surrogate als Bewertungsgegenstand bisheriger Studien zu Relevanzkriterien gelegt (Abschnitt 2.1.4), bevor eine Zusammenfassung über die Kriterien bei der Relevanzbewertung aus inhaltlicher Sicht erfolgt (Abschnitt 2.1.5). Der zweite Teil (Abschnitt 2.2) beschreibt die Methoden, die in bisherigen Studien zur Erforschung von Relevanzkriterien verwendet wurden. Dieser Teil ist jeweils in zwei Methoden der Datenerhebung und in zwei Methoden der Untersuchungsart gegliedert: Zunächst werden Befragungen (Abschnitt 2.2.1) und Beobachtungen (Abschnitt 2.2.2) beleuchtet, anschließend werden explorative Untersuchungsdesigns (Abschnitt 2.2.3) und schließlich experimentelle Untersuchungsdesigns (Abschnitt 2.2.4) vorgestellt. Daran schließt sich die Zusammenfassung der Methoden bisheriger Studien zur Erforschung von

Relevanzkriterien an (Abschnitt 2.2.5). Auf der Basis der Erkenntnisse der Literaturschau werden die Forschungslücken aufgezeigt und die Forschungsfragen abgeleitet (Abschnitt 2.3).

Im anschließenden **Kapitel 3** werden die Voraussetzungen zur experimentellen Erforschung von Relevanzkriterien herausgearbeitet, indem das informationswissenschaftliche Relevanzkonzept spezifiziert wird (Abschnitt 3.1). Zu der Spezifikation gehören die Erläuterung verschiedener Relevanzformen und der dieser Arbeit zugrunde gelegten Relevanzdefinition (Abschnitt 3.1.1), die Darstellung von Relevanz als ein multidimensionales Konstrukt (Abschnitt 3.1.2), und die Argumentation, den Prozess der Relevanzbewertung von Suchergebnissen als einen Prozess des Urteilens zu betrachten (Abschnitt 3.1.3). Des Weiteren werden in diesem Kapitel die Einflüsse, die im Bewertungsprozess von Bedeutung sind, identifiziert und mithilfe eines Modells veranschaulicht (Abschnitt 3.2). Diese Einflüsse lassen sich in drei Aspekte unterteilen – Elemente des Surrogats als potenzielle Relevanzmerkmale, subjektive Relevanzkriterien und Relevanzfaktoren: Zum einen stellen die Attribute eines Suchergebnisses die Elemente bzw. Merkmale dar, anhand derer informationssuchende Personen die Kriterien zur Bewertung bilden (Abschnitt 3.2.1); zum anderen wirken Einflussfaktoren als Kontext der Bewertung auf diesen Bewertungsprozess ein (Abschnitt 3.2.2). Das Kapitel endet mit der Zusammenfassung und Beantwortung der Forschungsfragen F1a und F1b (Abschnitt 3.3).

Den Hauptteil der vorliegenden Arbeit bildet **Kapitel 4** mit der Beschreibung der empirischen Studie zur Untersuchung des Einflusses von Popularitätsdaten auf die Relevanzbewertung von Suchergebnissen in akademischen Suchsystemen. Das Kapitel ist in insgesamt sechs Abschnitte gegliedert und beginnt mit der Entwicklung des experimentellen Untersuchungsdesigns (Abschnitt 4.1). Zu diesem gehören die Auswahl und Operationalisierung der unabhängigen Variablen (Abschnitt 4.1.1), die Bestimmung der abhängigen Variable (Abschnitt 4.1.2) sowie die Bildung der mit dem Experiment zu überprüfenden Hypothesen (Abschnitt 4.1.3). Ergänzend dazu wird der Umgang mit Drittvariablen, die als potenzielle Störvariablen mit den zu untersuchenden Variablen konfundieren könnten, erläutert (Abschnitt 4.1.4). Der Abschnitt zum Untersuchungsdesign schließt mit der ausführlichen Beschreibung des Versuchsaufbaus ab (Abschnitt 4.1.5).

Im Anschluss werden die Planung und Umsetzung der Datenerhebung ausführlich erläutert (Abschnitt 4.2). Hierzu zählen die Entwicklung der durch die Versuchspersonen zu bearbeitenden Aufgaben, die Beschreibungstexte zu Informationsbedürfnissen (Abschnitt 4.2.1) und zu bewertende Surrogate (Abschnitt 4.2.2) beinhalten. Die Daten wurden mithilfe eines Online-Fragebogens erhoben, dessen

Kern das eigentliche Experiment neben der Erfassung zusätzlicher Angaben der Stichprobe darstellt (Abschnitt 4.2.3). Schließlich wird die Berechnung der erforderlichen Stichprobengröße vorgestellt (Abschnitt 4.2.4) und das Vorgehen bei der Probandenakquise dargelegt (Abschnitt 4.2.5). Nach dem Abschnitt zur Datenerhebung werden die Aufbereitung (Abschnitt 4.3.1) und statistische Analyse der Daten (Abschnitt 4.3.2) beschrieben, bevor die Ergebnisse des Experiments vorgestellt werden (Abschnitt 4.4). Dieser Abschnitt beginnt mit der Betrachtung der Stichprobe (Abschnitt 4.4.1), an die sich die Vorstellung der Haupteffekte (Abschnitt 4.4.2) schließt. Das Hauptaugenmerk der Ergebnisdarstellung liegt in der Erläuterung der Interaktionseffekte zwischen den untersuchten unabhängigen Variablen (Abschnitt 4.4.3). Die Diskussion der Ergebnisse erfolgt zunächst im Kontext der Studienmethodik (Abschnitt 4.5), anhand derer auch die Grenzen der Studie aufgezeigt werden (Abschnitt 4.6).

Kapitel 5 beinhaltet die Schlussbetrachtungen dieser Arbeit. Diese beginnen mit der Beantwortung der Forschungsfragen F1, F2 und F3 (Abschnitt 5.1). Anschließend werden die Ergebnisse der Arbeit im Kontext der Gesamtmethodik reflektiert (Abschnitt 5.2), bevor ein Ausblick auf künftige Forschung erfolgt (Abschnitt 5.3).

Dem Anhang im elektronischen Zusatzmaterial zu entnehmen sind eine Übersicht der analysierten Studien zu Relevanzkriterien (Anhang 1), der Fragebogen zur Datenerhebung (Anhang 2), die SPSS-Syntax der statistischen Mehrebenenanalyse (Anhang 3) und weitere Ergebnisse der Datenauswertung (Anhang 4).

Stand der Forschung

2

Die Darstellung des Forschungsstands in diesem Kapitel erfolgt aus zwei Perspektiven – einer inhaltlichen und einer methodischen Perspektive auf die Kriterien, die informationssuchende Personen bei der Relevanzbewertung von Suchergebnissen anwenden. Der erste Teil dieses Kapitels (Abschnitt 2.1) widmet sich den Kriterien aus inhaltlicher Sicht. Dazu wird einleitend auf das Problem einer fehlenden definitorischen Abgrenzung der in der Literatur verwendeten Begriffe „Kriterien", „Faktoren" und „Merkmale" eingegangen und aufgezeigt, welche Konsequenzen sich diesbezüglich für die Auswahl und Betrachtung der für die Erläuterungen der Relevanzkriterien herangezogenen Studien in den nachfolgenden Abschnitten 2.1.1, 2.1.2 und 2.1.3 ergeben. Anschließend wird der Fokus auf Surrogate als zu bewertendes Informationsobjekt in Studien zu Relevanzkriterien gelegt und deren besonderer Stellenwert bei der Relevanzbewertung verdeutlicht (Abschnitt 2.1.4). In Abschnitt 2.1.5 erfolgt eine Zusammenfassung der Erkenntnisse über die Kriterien bei der Relevanzbewertung, anhand derer die mit dieser Arbeit adressierte Forschungslücke deutlich wird.

Die separate Betrachtung der Methoden zur Erforschung von Relevanzkriterien im zweiten Teil dieses Kapitels (Abschnitt 2.2) erfolgt vor dem Hintergrund der besonderen Herausforderung, ein komplexes Konzept wie Relevanz zu messen. Buckland (2017) betont, dass Relevanz keine direkt messbare, physische Einheit darstellt wie beispielsweise Messeinheiten in den Naturwissenschaften, sondern aufgrund der subjektiven und dynamischen Natur von Relevanz diese im direkten Zusammenhang mit menschlicher kognitiver Aktivität steht.

Ergänzende Information Die elektronische Version dieses Kapitels enthält Zusatzmaterial, auf das über folgenden Link zugegriffen werden kann https://doi.org/10.1007/978-3-658-37512-6_2.

C. Behnert, *Popularität und Relevanz in der Suche*,
https://doi.org/10.1007/978-3-658-37512-6_2

Damit ein Dokument als relevant gilt, „[muss es] useful to an actual human being's mental activity [sein]" (Buckland, 2017, S. 161). Es ist davon auszugehen, dass mentale Aktivität eng verknüpft ist mit menschlichem Verhalten. Um menschliches Erleben und Verhalten zu erforschen, werden in Disziplinen, die sozialwissenschaftliche Methoden anwenden, wie der Psychologie, häufig Experimente durchgeführt. Mithilfe von Experimenten sollen kausale Zusammenhänge zwischen einer Ursache (Stimulus) und deren Wirkung (beobachteter Effekt) hergestellt werden. Dabei werden die als ursächlich auf eine Wirkung vermuteten Faktoren manipuliert und können systematisch und isoliert untersucht werden.

Die Bibliotheks- und Informationswissenschaft greift traditionell auf Methoden und Verfahren zur Erhebung und Analyse von Daten aus anderen Fachdisziplinen zurück. Dies sind vor allem sozialwissenschaftliche Methoden und Verfahren (z. B. bei Umfragen, Nutzerstudien) sowie Anwendungen aus dem Bereich Informatik (z. B. Programmierung von Software-Tools). Um Effekte auf tatsächliches menschliches Verhalten zu untersuchen, werden auch in der IIR-Forschung Experimente durchgeführt (Kelly, 2009; Kelly & Cresenzi, 2016). Studien, in denen verschiedene Faktoren, die eine Rolle dabei spielen, wie Menschen die Relevanz von Informationen beurteilen, untersucht werden, bezeichnet Saracevic (2007b) als *relevance behavior studies* und hebt damit den Verhaltensaspekt in diesen Studien hervor. In einigen Publikationen zu Relevanz und Relevanzkriterien werden zudem konkret die Begriffe Relevanzverhalten (*relevance behavior/behaviour*) verwendet (z. B. Balatsoukas et al., 2010; Borlund, 2003b; Cook, 1971; Ruthven, 2014; Saracevic, 2007b, 1996; Scholer et al., 2013; Wang, 2011). Vor diesem Hintergrund erscheint es sinnvoll, zu analysieren, welche Methoden bisherige Studien zur Erforschung von Relevanzkriterien verwendeten und insbesondere, ob unter ihnen ebenfalls Experimente sind.

Am Ende des Kapitels werden die sich anhand der Betrachtung des Forschungsstands eröffneten Forschungslücken zusammengefasst und die sich daraus ableitenden Forschungsfragen, die mit der hier vorgestellten Forschung untersucht wurden, formuliert (Abschnitt 2.3).

2.1 Kriterien bei der Relevanzbewertung

Ein tiefgehendes Verständnis der Kriterien, die der subjektiven Relevanzbewertung von Informationsobjekten zugrunde liegen, ist notwendig, um zu erkennen, wodurch sich ein Dokument als mehr oder weniger relevant zu einer Suchanfrage oder dem Informationsbedürfnis einer Person gegenüber einem anderen Dokument auszeichnet. Jahrzehntelange Relevanzforschung hat gezeigt, dass es

einfacher scheint zu beschreiben, was ein irrelevantes Dokument ausmacht, als zu definieren, was ein relevantes Dokument ist (Hjørland, 2000). Aufgrund des hohen Maßes an Subjektivität, Dynamik und Multidimensionalität, welches das informationswissenschaftliche Relevanzkonzept kennzeichnet (Mizzaro, 1997) und des Fehlens seiner allgemein gültigen Definition (Saracevic, 2016b), lässt sich der Begriff des Relevanzkriteriums offenbar ebenso wenig eindeutig und universell definieren. In den Studien zu Relevanzkriterien sind unterschiedliche Definitionen des Begriffs Relevanzkriterien (*relevance criteria*) zu finden, zum Beispiel:

- „[C]riteria for relevance judgments […] are reasons underlying human assessments“ (Schamber, 1994, S. 13).
- „A relevance criterion can be defined as the parameter or value by which users determine the relevance of a retrieved object at a certain point in time“ (Balatsoukas & Ruthven, 2012, S. 1728).
- „Relevance criteria are defined as the factors or reasons that influence users' relevance judgments“ (Schamber & Bateman, 1999, S. 382).

Diese exemplarisch ausgewählten Zitate bieten neben Gründen und Parametern oder Werten auch den Begriff Faktoren an; in einer weiteren Erläuterung im Zusammenhang mit der Suche nach Gesundheitsinformationen ist von Regeln zur Auswahl von Quellen die Rede: „Criteria are rules by which users select a source“ (Zhang, 2014, S. 915). Für eine klare Definition des Begriffs Kriterium kommt erschwerend hinzu, dass neben den Relevanzkriterien in der Literatur ebenfalls die Begriffe Relevanzfaktoren (*relevance factors*, *factors of relevance*) und Relevanzmerkmale (*relevance clues*, *clues to relevance*, *relevance cues*) auftauchen. Beispielsweise findet sich bei Saracevic (2016b) eine unglückliche Vermischung von Attributen und Kriterien: „[C]lues research aims to uncover and classify attributes or criteria that users concentrate on while making relevance inferences“ (2016b, S. 50). Aufgrund von terminologischen Differenzen lassen sich die Ergebnisse aus unterschiedlichen empirischen Studien zu Relevanzkriterien nur schwer einordnen und miteinander vergleichen, wie bereits Bales & Wang (2006) und Wang (2010) beschreiben, wodurch insbesondere systematische Literaturschauen oder Metaanalysen anfällig für Fehlinterpretationen oder das Ziehen falscher Schlüsse sind. So listen manche Studien beispielsweise Aktualität als Einflussfaktor auf (z. B. Schamber, 1994), während andere Aktualität als Kriterium ausweisen (z. B. Barry, 1994). In seiner Synthese über jahrzehntelange informationswissenschaftliche Relevanzforschung stellt Saracevic fest:

> Criteria, language, measures, and methods used in […] studies [that contain data directly addressing relevance] were not standardized and they varied widely. In that sense, although no study was an island, each study was done more or less on its own. Thus, the results are only cautiously comparable. (2016b, S. 57)

Das Definitionsproblem offenbart sich insbesondere in der Übersicht der 80 Faktoren in 6 Kategorien (Tabelle 2.1), die basierend auf einer umfassenden Literaturschau zusammengetragen wurden (Schamber, 1994, S. 11). Die Kategorien *Judgment conditions* und *Choice of scale* zeigen, dass es sich um Faktoren, die im Forschungskontext zur Erhebung expliziter Relevanzbewertungen beobachtet wurden, handelt. Auffällig sind Begriffe wie *Pertinence*, *Usefulness* und *Difficulty level* als Faktoren der Kategorie *Documents*, die im Gegensatz zu *Aboutness* in derselben Kategorie nicht unveränderbar mit dem Dokument verbunden sind, sondern veränderbare Konzepte bezeichnen, deren Werte bzw. Urteile abhängig von der Person sind, durch die die Bewertung vorgenommen wird. Auch ob der Inhalt des Dokuments interessant ist (*Interesting content*), kann ohne den Bezug zum Subjekt nicht beurteilt werden. Die als Faktoren aufgeführten Begriffe in dieser Kategorie *Documents* vermischen objektiv erfassbare Eigenschaften des Dokuments wie *Aboutness* oder *Recency* mit den subjektiv zu bestimmenden Eigenschaften, die einem Dokument ausschließlich in Abhängigkeit mit dem Kontext der bewertenden Person und ihrem subjektiven Informationsbedürfnis zugeschrieben werden können. Des Weiteren sind manche Begriffe unklar in ihrer genauen Bedeutung, zum Beispiel in Hinblick auf den Unterschied zwischen *Novelty* und *Recency;* unscharf sind Begriffe wie *Cognitive style* in der Kategorie *Judges* und *Authorship* in der Kategorie *Documents.* Aus diesen Gründen lassen die gelisteten Faktoren keine eindeutige Definition und Abgrenzung gegenüber Kriterien zu, sondern erlauben einen relativ weiten Interpretationsspielraum.

Tabelle 2.1
Relevanzfaktoren aus dem Literaturbericht von Schamber (1994, S. 11)

JUDGES	Diversity of content (1)
Biases (1)	Importance (3)
Cognitive style (1)	Informativeness (3)
Concept of relevance (1,2)	Interesting content (3)
Error preference (1)	Level of condensation (1,2)
Expectations regarding distribution (1)	Logical relevance (3)
Formal education (2)	Novelty (3)
Intelligence (1)	Pertinence (3)
Judging experience (1)	Publication source (3)
Judgment attitude (1)	Recency (3)
Knowledge/experience (1,2)	Scientific “hardness” (1,2)
Professional involvement (2)	Specificity / amount of information (1,2)
Research stage (2)	Style (1,3)
Use orientation (1,2)	Subject matter (1)
Vigilance level (1)	Textual attributes (1)
JUDGMENT CONDITIONS	Usefulness (2,3)
Breadth of document set (1)	INFORMATION SYSTEM
Definition of relevance (1,2)	Access (item identification) (4)
Order of presentation (1)	Access (subject description) (4)
Size of document set (1)	Access (subject summary) (4)
Social pressure toward convergence (1)	Accuracy (data transfer) (4)
Specification of the task (1,2)	Browsability (4)
Time for judging (1)	Comprehensiveness (coverage) (4)
Use of control judgments (1)	Convenience of location (3)
CHOICE OF SCALE	Convenience of hours (3)
Availability of anchors (1)	Cost saving (4)
Ease of use (1,2)	Currency (updating) (4)
Kind of response required (1,2)	Ease of detection of relevance (3)
Number of rating categories (1)	Effort expended (3)
Type of scale (1,2)	Flexibility (dynamic interaction) (4)
REQUESTS	Formatting (scannability) (4)
Diversity of content (1)	Interfacing (help, orientation) (4)
Difficulty level (1)	Links to external sources (4)
Functional ambiguity (1)	Ordering (subject matter) (4)
Specificity / amount of information (1,2)	Physical accessibility (4)

(Fortsetzung)

Tabelle 2.1 (Fortsetzung)

DOCUMENTS	Precision of subject output (4)
Aboutness (3)	Reliability (consistency (4)
Accuracy (truth) (3)	Response speed (4)
Aesthetic value (3)	Selectivity (input choices) (4)
Authorship (3)	Simplicity (clarity) (4)
Credibility (3)	Time spent (3)
Difficulty level (1)	

Basiert auf (1) Cuadra & Katter, (2) Rees & Schultz, (3) Cooper (1971; 1973), (4) Taylor, beinhaltet alle Faktoren, die vorgeschlagen wurden von Cuadra und Katter und ausgewählte Faktoren, vorgeschlagen von anderen:
(1) Cuadra, C. A.; Katter, R. 1967. Experimental Studies of Relevance Judgments Final Report. Volume I: Project Summary. Santa Monica, CA: System Development Corp.; 1967. 129p. ITM-3520/001/00). NTJS: PB 175518.
(2) Rees, A. M.; Schultz, D. G. 1967. A Field Experimental Approach to the Study of Relevance Assessments in Relation to Document Searching: Final Report: Volume I. Cleveland, OH: Case Western Reserve University, School of Library Science, Center for Documentation and Communication Research; 1967. 30Sp. NTIS: PB 176 080.
(3) Cooper, W. 1971. A Definition of Relevance for Information Retrieval. Information Storage and Retrieval. 1971 June; 7(1); 19–37. ISSN: 0020–0271. Und Cooper, W. 1973. On Selecting a Measure of Retrieval Effectiveness. Journal of the American Society for Information Science. 1973 March/April; 24(2): 87–100. ISSN: 0002–8231.
(4) Taylor, R. 1986. Value-Added Processes in Information Systems. Norwood, NJ: Ablex Publishing Corp.; 1986. 257p. ISBN: 0–89391-273–5.

Saracevic (2016b) fasst in seiner Literaturschau die Erkenntnisse aus jahrzehntelanger informationswissenschaftlicher Relevanzforschung zusammen und beschreibt unter anderem die Faktoren, die Effekte auf Relevanzbewertungen ausüben, als: (a) Faktoren im Zusammenhang mit den Jurorinnen und Juroren, wie Erfahrung mit thematisch-relevanten Fragen, Kenntnis über und Interesse an einem Thema, spezielles Fachwissen, Sprache; und (b) Faktoren im Zusammenhang mit den Relevanzbewertungen, wie die Position des Treffers in der Ergebnisliste, Design und Usability des IR-Systems, Art der Suchaufgabe/-ergebnisse (Schwierigkeitsgrad, Informationen) und die Art der Bewertung (Skala) bei der Erhebung expliziter Relevanzbewertungen (Saracevic, 2016b, S. 77 ff.). Diese

Einteilung von Faktoren stellt bereits eine konkrete Abgrenzung des Begriffs Relevanzfaktor von dem Begriff Relevanzkriterium dar und erleichtert die Identifizierung der Studien, die Relevanzkriterien erforschen, dadurch, dass diese nicht auf Einflussfaktoren, sondern auf andere, subjektiv ableitbare Parameter fokussiert sind.

Die Erkenntnisse der nachfolgenden Abschnitte beruhen zum einen auf den bedeutenden und in der informationswissenschaftlichen Literatur zu Relevanz vielzitierten Arbeiten von Stefano Mizzaro (z. B. Mizzaro, 1997) und Tefko Saracevic (z. B. Saracevic, 2016b), zum anderen auf den Ergebnissen von informationswissenschaftlichen, empirischen Studien zu Relevanzkriterien, die vordergründig für die Auseinandersetzung mit den methodischen Aspekten in Abschnitt 2.2 mithilfe der *Chaining*-Methode[1] identifiziert und anhand vorab festgelegter Kriterien[2] ausgewählt wurden. Für die Betrachtung von Relevanzkriterien aus inhaltlicher Sicht wurden neben diesen analysierten Studien weitere herangezogen, die für die Analyse in Hinblick auf methodische Aspekte aufgrund der Auswahlkriterien nicht infrage kamen. Aus diesem Grund werden in den nachfolgenden Abschnitten auch Quellen zitiert, die in Abschnitt 2.2 nicht berücksichtigt werden.

Die Ergebnisse der Studien zu Relevanzkriterien lassen sich chronologisch anhand ihrer Veröffentlichung in allgemeine Kriterien (Abschnitt 2.1.1) und Kriterien im Kontext der Websuche (Abschnitt 2.1.2) gruppieren, da in frühen Studien, insbesondere in denen der 1990er Jahre, anhand von offenen Forschungsfragen sehr allgemein untersucht wurde, welche Kriterien bei der Relevanzbewertung auftreten; später wuchs mit dem Aufkommen elektronisch zugänglicher Inhalte über das WWW das Forschungsinteresse nach den Kriterien bei der Websuche, die insbesondere durch Autorität und Glaubwürdigkeit gekennzeichnet sind, unabhängig von dem Kontext der gesuchten Informationen (Abschnitt 2.1.2.1). In diesem Zusammenhang kann Popularität als Indikator für Qualität eine besondere Bedeutung zugesprochen werden. Zum Zeitpunkt der Erfassung dieses Forschungsstands gibt es allerdings keine Studien zu Relevanzkriterien in textbasierten IR-Systemen, die Popularität als Kriterium bei der Bewertung explizit berücksichtigen. Aus diesem Grund beschränkt sich Abschnitt 2.1.2.2 auf die Erläuterung des Konzepts der Popularität und gibt den theoretischen Rahmen für die Betrachtung der Studien zu Relevanzkriterien in

[1] Die Chaining-Methode bezeichnet eine Strategie bei der Suche nach Dokumenten, bei der zitierte Quellen in einem Werk verfolgt (*backward chaining*) sowie die dieses Werk zitierenden Publikationen identifiziert (*forward chaining*) werden (Ellis, 1989).

[2] Das konkrete Vorgehen bei der Auswahl der Studien wird in Abschnitt 2.2 beschrieben.

den nachfolgenden Abschnitten 2.1.3 und 2.1.4 vor: Die Konzepte Autorität und Popularität lassen sich auf die Relevanzbewertung in akademischen Suchsystemen (z. B. Websuchmaschinen mit wissenschaftlichen Inhalten, fachdisziplinspezifische oder fachübergreifende Datenbanken) übertragen (Abschnitt 2.1.3).

Anhand welcher Kriterien Nutzerinnen und Nutzer bewusst oder unbewusst ein Informationsobjekt bewerten, sollte stets im Zusammenhang mit den vorhandenen Attributen oder Eigenschaften des zu bewertenden Objektes untersucht werden. So zeigen Studien, dass der Titel und eine Zusammenfassung (Abstract) die wichtigsten Attribute für das Ableiten von thematischer Relevanz sind; Kriterien und deren Gewichtung bei der Relevanzbewertung sind nach der Art (und somit auch des Umfangs) des Bewertungsgegenstands nicht immer dieselben, wenngleich ähnlich (Saracevic, 2016b, S. 58). Daher ist es ebenfalls erforderlich zu präzisieren, ob die Relevanzbewertung auf der Bewertung des Surrogats oder des Volltexts beruht. Watson (2014) nimmt eine Einteilung von Relevanzkriterien in Kriterien vor dem Zugang (*pre-access criteria of relevance*) und Kriterien nach dem Zugang (*post-access criteria of relevance*) zu einem Volltext-Dokument vor. Diese Art der Unterscheidung impliziert jedoch, dass eine klare Trennung der Kriterien vor und nach dem Zugriff auf den Volltext ohne eine mögliche Überschneidung vorgenommen werden kann.

Eine alternative Unterscheidungsart, die ebenfalls den Gegenstand und somit auch den Zeitpunkt der Bewertung berücksichtigt, stellt die Einteilung der Art der Relevanzbewertungen in *predictive judgments* und *evaluative judgments*[3] dar: Die Relevanzbewertung erfolgt entweder anhand des Surrogates und somit vor dem Zugang zum Volltext (*predictive judgment*) oder auf der Basis des Dokumenteninhalts nach dem Zugang zum Volltext (*evaluative judgment*) (Rieh, 2002). Nicht alle Studien zu Relevanzkriterien nehmen eine Unterscheidung zwischen *predictive judgments* und *evaluative judgments* vor, wie es bei beispielsweise bei Rieh (2002) der Fall ist. In wenigen Studien wird die Tatsache berücksichtigt, dass der Zeitpunkt der Bewertung (vor oder nach dem Zugang zum Volltext) zugleich Aufschluss über das Aussehen und den Umfang der Grundlage der Relevanzbewertung und somit über die Grundlage für die Ableitung oder Bildung von Relevanzkriterien gibt. Vor diesem Hintergrund sind Surrogate als Bewertungsgrundlage von besonderem Interesse (Abschnitt 2.1.4), denn diese enthalten mitunter Informationen, die beispielsweise auf die Popularität des repräsentierten

[3] Eine wörtliche Übersetzung der Begriffe *predictive* (prädiktiv) und *evaluative* (evaluierend) im Kontext von Relevanzbewertungen erscheint unpassend; daher werden fortan die englischen Begriffe verwendet oder umschrieben, wie „Surrogatbewertung“ für *predictive*, „Bewertung des Volltexts“ für *evaluative judgment*.

Werkes hindeuten sollen und somit bei der Bildung von *predictive judgments* eine Rolle spielen können.

2.1.1 Allgemeine Kriterien

Thematische Relevanz gilt übereinstimmend als Grundbedingung für Relevanz (Cosijn & Ingwersen, 2000; Greisdorf, 2003; A. R. Taylor et al., 2007; Wang & Soergel, 1998; Xu & Yin, 2008) und als Basis für diverse weitere Kriterien, die auf ebendieser aufbauen:

> All relevance judgments start with topically relevant materials (which is an appropriate first step of systems), but then diverse criteria come into play operating dynamically in a process in which certain citations are rejected or accepted on one or more criteria (Froehlich, 1994, S. 129).

Thematische Relevanz (*topical relevance, subject relevance*) kennzeichnet die Beziehung zwischen dem Thema einer Suchanfrage und dem Thema des gefundenen Informationsobjekts (Saracevic, 1996). Dabei sind die Begriffe nicht zu verwechseln mit der Bezeichnung *subjective relevance*, die auf die subjektive Sichtweise von Relevanz abzielt. Hjørland (2010) verortet *topicality* eindeutig auf der Seite der Nutzer (*user's view*) und nicht auf der Seite des Systems (*system's view*).

Das Konzept von *Aboutness*[4] ist in diesem Zusammenhang von zentraler Bedeutung, weil es aus systemseitiger Sicht gleichgesetzt wird mit objektiver Relevanz – ein Dokument ist objektiv relevant zu einer Suchanfrage, wenn es von dem Inhalt der Suchanfrage handelt –, obwohl der Begriff Relevanz in diesem Zusammenhang unangemessen ist (Harter, 1992).

Aboutness und Relevanz sind zwar verwandte Konzepte, unterscheiden sich aber dahingehend, dass sich Aboutness auf Fachgebiete oder Themen bezieht,

[4] Für den englischsprachigen Begriff *Aboutness* gibt es keine deutschsprachige Übersetzung. Lt. Eintrag im Oxford-Wörterbuch liegt dem Begriff eine philosophische Bedeutung zugrunde: „The quality or fact of relating to or being about something; Philosophy (of a mental state, symbol, representation, etc.) the property of being about something (existent or non-existent)." (https://en.oxforddictionaries.com/definition/aboutness, letzter Zugriff am 05.05.2017). Im Kontext des Indexierens bezeichnet der Begriff Aboutness die Fähigkeit zu erkennen, wovon der Inhalt eines Dokuments handelt (Maron, 1977). Für weiterführende Literatur zu Aboutness als informationswissenschaftliches Konzept siehe beispielsweise Bruza et al. (2000), Hjørland (2001), O'Neill et al. (2017).

Relevanz auf ein (Informations-)Problem – „The fundamental notion in organization of information is aboutness, while the fundamental notion in searching is relevance“ (Saracevic, 2012, S. 58). Aboutness ist im Gegensatz zu Relevanz nicht dynamisch und kann für ein Dokument unabhängig von einer Suchanfrage oder einem Informationsbedürfnis bestimmt werden. Das bedeutet: „Documents can, however, have the same subject (or the same aboutness) without having the same relevance“ (Hjørland, 2001, S. 777).

Vor diesem Hintergrund führt Borlund (2003b) den Begriff *intellectual topicality* ein, der sich von der als algorithmisch oder objektiv bestimmten Relevanz aus Systemsicht von der intellektuell bestimmten thematischen Relevanz durch den Menschen klar abgrenzt. In ihrer Studie greifen Xu & Chen (2006) auf eine ähnliche Unterscheidung zurück und definieren *topicality* nicht als thematische Relevanz, sondern als Eigenschaft des Dokuments und somit im Sinne von Aboutness:

> In this study, topicality is regarded as a document attribute rather than relevance itself; the term relevance refers to the portion of the relevance continuum beyond topicality; it encompasses both cognitive and situational relevance. We define it as the perceived cognitive and pragmatic impact of the content of a document in relation to the user's problem at hand. (Xu & Chen, 2006, S. 962)

Studien zu Relevanzkriterien zeigen, dass Menschen für ihre Relevanzbewertungen eine Vielzahl an Kriterien identifizieren, die über die thematische Relevanz hinausgehen. Saracevic (2016b) gibt einen Überblick über 21 empirische Studien zu Relevanzkriterien, die er in sieben Kategorien und zwei Gruppen zusammenfasst (Tabelle 2.2): *Content, Object, Validity* im Zusammenhang mit den Eigenschaften des Informationsobjekts; *Usefulness or situational match, Cognitive match, Affective match, Belief match* im Zusammenhang mit den Eigenschaften der informationssuchenden Person. Dabei fügt er unter anderem *topic, quality, depth, scope* als Relevanzkriterien der Kategorie Content hinzu. Qualität als eigenständiges Relevanzkriterium ist allerdings sehr zu hinterfragen, weil es wesentlicher breiter gefasst ist als beispielsweise *scope* und dies zu der Frage führt, anhand welcher Kriterien wiederum die Qualität des Informationsobjekts beurteilt wird. Xie & Benoit (2013) ordnen hingegen *scope* und *depth* der thematischen Relevanz zu.

Tabelle 2.2 Relevanzkriterien aus Saracevic (2016b, S. 57 ff.)

Kategorie	Relevanzkriterien	Interaktion zwischen…
Content	Topic, quality, depth, scope, currency, treatment, clarity	Information (or object) characteristics
Object	Characteristics of information objects, e.g., type, organization, representation, format, availability, accessibility, costs	
Validity	Accuracy of information provided, authority, trustworthiness of sources, verifiability, reliability	
Usefulness or situational match	Appropriateness to situation, or tasks, usability, urgency; value in use	Individual (or human) characteristics
Cognitive match	Understanding, novelty, mental effort. Link to previous knowledge	
Affective match	Emotional responses to information, fun, frustration, uncertainty	
Belief match	Personal credence given to information, confidence	

Des Weiteren betont Saracevic (2016b) die Interaktion zwischen den jeweiligen Kriterien, d. h. Relevanzkriterien können nicht getrennt voneinander betrachtet werden. Die von ihm vorgestellten Studien wurden zwischen 1990 und 2015 veröffentlicht, unter ihnen sind die bedeutenden Arbeiten von Barry und Schamber (Barry, 1994, 1998; Barry & Schamber, 1998; Schamber, 1991; Schamber & Bateman, 1999), die einen wichtigen Grundstein für nachfolgende Studien zu Relevanzkriterien lieferten.

Barry & Schamber (1998) verglichen die Ergebnisse ihrer empirischen Studien zu den Kriterien, nach denen Testpersonen die Relevanz eines Dokuments bewerteten, miteinander. Auf Basis von Inhaltsanalysen der Interviews von 18 (Barry, 1994) bzw. 30 Testpersonen (Schamber, 1991) entwickelten sie jeweils 23 bzw. 10 Kategorien von Kriterien und verglichen diese nach der Häufigkeit ihrer Nennung

in den Interviews. Sie gelangten zu der Schlussfolgerung, dass Relevanzbewertungen abhängig sind von der individuellen Wahrnehmung der jeweiligen Person bezüglich ihres Informationsproblems und ihrer Informationsumgebung. Dabei sind etliche Relevanzkriterien zusätzlich zum Inhalt des Informationsobjekts von Bedeutung, wie beispielsweise *Validität, Aktualität, Verfügbarkeit* und *Vertrauenswürdigkeit der Informationsquelle* (Barry & Schamber, 1998). Als besonders wichtiges Ergebnis aus diesem Studienvergleich ist die Tatsache zu nennen, dass Barry und Schamber zwei verschiedene Nutzergruppen in unterschiedlichen Kontexten (Studierende im akademischen Kontext bzw. Personen im berufsbezogenen Kontext) untersuchten und sich herausstellte, dass sich die verwendeten Relevanzkriterien bei den Gruppen überschneiden. Dies lässt die Schlussfolgerung zu, dass es eine begrenzte Menge an Kriterien (*a finite range of criteria*) gibt, die universell wirken und je nach Kontext entsprechend angepasst werden (Barry & Schamber, 1998, S. 234).

Insgesamt benennen die bisherigen Studien sehr viele verschiedene Kriterien, die zusammengenommen eine eher unübersichtliche Menge darstellen. Anhand der Zahl der verwendeten Relevanzkriterien wird erneut die Komplexität des Relevanzbewertungsprozesses deutlich (Beresi, 2011). Um dieser Vielzahl systematisch zu begegnen, entwickelten Xu & Chen (2006) ein Fünf-Faktoren-Modell von Relevanz (*five-factor model of relevance*), bei dem es sich jedoch vielmehr um ein Kriterien-Modell handelt. Diese fünf Kriterien bezeichnen die Autoren als Schlüsselkriterien: *topicality, novelty, reliability, understandability, scope*. Zu beachten ist hier, dass *topicality* und *scope* nicht gleichgesetzt, sondern als sich ergänzende Kriterien aufgezählt werden.

Die in diesem Abschnitt vorgestellten Kriterien können als universell geltende Kriterien betrachtet werden. Allerdings kommt manchen Kriterien eine besondere Bedeutung bei der Bewertung von Suchergebnissen während der Websuche zu, daher werden sie nachfolgend in einem eigenen Abschnitt näher erläutert.

2.1.2 Kriterien im Kontext der Websuche

Bei den Studien, die nach der Jahrtausendwende veröffentlicht wurden, fällt auf, dass sich der Forschungsfokus von der Identifizierung allgemeiner Relevanzkriterien entfernt und auf spezielle oder einzelne Kriterien in einem bestimmten Kontext richtet. Manche Studien zu Relevanzkriterien berücksichtigen gezielt den Kontext der Websuche, wie beispielsweise die Studie von Tombros, Ruthven, & Jose (2003, 2005), für die 24 Testpersonen die Nützlichkeit (*usefulness*) von Webseiten bewerteten, und neben dem Inhalt der Webseiten deren Struktur

(z. B. Layout) als wichtiges Kriterium für die Auswahl ermittelt wurde. Savolainen & Kari (2006) untersuchten Kriterien für die Auswahl von Hyperlinks und Webseiten, indem sie Videoaufnahmen von 9 Testpersonen bei der Websuche und unter Nutzung der Think-aloud-Methode[5] auswerteten. Diese Studien konnten zeigen, dass auch für das Aufrufen von Webseiten thematische Relevanz als Hauptkriterium gesehen werden kann.

Rieh & Belkin (1998) argumentieren, dass durch die enorme und vor allem dynamische Menge an Informationsobjekten ein vollständiger Recall im Web nicht möglich ist, sodass sehr viele Informationen gefiltert[6] werden müssen, die oft keiner Qualitätskontrolle (wie beispielsweise dem Peer Review bei wissenschaftlichen Publikationen) unterzogen wurden. Daher gelten neben thematischer Relevanz insbesondere im Kontext der Websuche die Kriterien der Qualität und Autorität. Rieh & Belkin (1998) wollten herausfinden, wie Menschen die Informationsqualität und Autorität von Informationen im Web bewerten und ob sie im Web andere Kriterien dabei anwenden als in traditionellen IR-Systemen. Auf der Basis von Interviews mit 14 Wissenschaftlerinnen und Wissenschaftlern identifizierten sie Autorität als zugrundeliegendes Konzept für die Glaubwürdigkeit einer Quelle im Web.

Da die Beurteilung von Autorität, Vertrauenswürdigkeit und Glaubwürdigkeit ausschließlich auf den individuellen Erfahrungen, dem Wissensstand und den persönlichen Überzeugungen der informationssuchenden Person beruhen (Rieh, 2009), stellen sie ebenfalls ausschließlich subjektive Relevanzkriterien dar. Diese werden in dem nachfolgenden Abschnitt 2.1.2.1 näher betrachtet. Daran anschließend wird in Abschnitt 2.1.2.2 das Konzept der Popularität beleuchtet, welches insbesondere bei der Websuche einen einflussreichen Faktor darstellt.

2.1.2.1 Glaubwürdigkeit und Autorität

Rieh (2009) definiert Glaubwürdigkeit (*credibility*) als „people's assessment of whether information is trustworthy based on their own expertise and knowledge"

[5] Bei dieser Methode werden die Testpersonen gebeten, ihre Gedanken und Handlungen während der Bearbeitung einer an sie gestellten Aufgabe simultan zu verbalisieren. Diese Methode der Datenerhebung wird in Abschnitt 2.2.1 im Zusammenhang mit den Methoden zur Erforschung von Relevanzkriterien näher beschrieben.

[6] Im Prinzip ergibt sich eine Art der Filterung bereits durch die Anzeige der Treffer innerhalb einer Suchergebnisseite, die von dem Ranking-Algorithmus bestimmt wird (z. B. mittels textstatistischer Verfahren in Verbindung mit nutzerbasierten Signalen). Auf diese systemseitige Auswahl von Quellen für die Anzeige der Treffer auf den vorderen Positionen im Gegensatz zu Treffern auf den hinteren Positionen wird mit Blick auf das Ziel dieser Arbeit nicht näher eingegangen.

(S. 1338). Xu & Chen (2006) betonen, dass Glaubwürdigkeit (*credibility*) nicht gleichzusetzen ist mit Zuverlässigkeit (*reliability*) – „the credibility of the source can be regarded as an external cue of document reliability […], we define reliability as the degree to which the content of a retrieved document is perceived to be true, accurate, or believable" (Xu & Chen, 2006, S. 964).[7]

Wathen & Burkell (2002) beleuchten in ihrem Literaturüberblick die Faktoren Glaubwürdigkeit und Vertrauenswürdigkeit (*credibility*[8]) von Informationsobjekten während der Websuche und betrachten Vertrauenswürdigkeit im Zusammenhang mit kognitiver Autorität (*cognitive authority*).

Das Konzept der kognitiven Autorität in der Informationswissenschaft wurde geprägt durch Patrick Wilson (1983)[9]. Er beschreibt, dass Menschen sich auf zwei Arten Wissen aneignen – einerseits aus der eigenen Erfahrung, andererseits aus dem durch andere Personen erfahrenen Wissen, was er als „second-hand knowledge" bezeichnet. Dabei gelten nur die Personen als glaubwürdig, deren Aussagen als wahr bzw. richtig anerkannt werden. Wenn diese Personen durch ihre Aussagen das Denken und Handeln anderer Personen beeinflussen, handelt es sich um kognitive Autoritäten: „The person whom I recognize as having cognitive authority is one whom I think should be allowed to have an influence on my thinking" (P. Wilson, 1983, S. 14). Olaisen (1990) ergänzt: „Others who are not cognitive authorities may also influence me. The difference between them and the cognitive authorities is that I recognise the latter's influence as proper and the former's as not proper" (Olaisen, 1990, S. 94).

Kognitive Autorität beinhaltet im Kern die beiden Konzepte Vertrauenswürdigkeit und Expertise bzw. Kompetenz (Rieh & Danielson, 2007, S. 312). Je höher der (wahrgenommene) Grad der Expertise einer Person (z. B. eines Autors), desto höher ist ihr Ansehen, was wiederum eine höhere Qualität der von dieser Person stammenden Information (z. B. in einem wissenschaftlichen Artikel) impliziert.

---

[7] Obwohl an dieser Stelle ein weiterer Begriff für glaubwürdig (*believable*) auftaucht, wird zwischen den beiden Konzepten *believability* und *credibility* im Rahmen dieser Arbeit nicht weiter unterschieden.

[8] Das Oxford Advanced Learners Dictionary definiert *credibility* als „the quality that somebody/something has that makes people believe or trust them" (https://www.oxfordlearnersdictionaries.com/definition/english/credibility, letzter Zugriff: 25.06.2020). Der Duden bietet zwar Kredibilität als deutschsprachigen Begriff für Glaubwürdigkeit an, verweist aber auf dessen veraltete Verwendung im österreichischen und schweizerischen Sprachgebrauch (https://www.duden.de/rechtschreibung/Kredibilitaet; letzter Zugriff: 25.06.2020). In dieser Arbeit wird für *credibility* der Begriff Glaubwürdigkeit verwendet.

[9] Eine kritische Auseinandersetzung mit dem Werk von Patrick Wilson bietet der Beitrag von White (2019) in der ISKO Encyclopedia of Knowledge Organization.

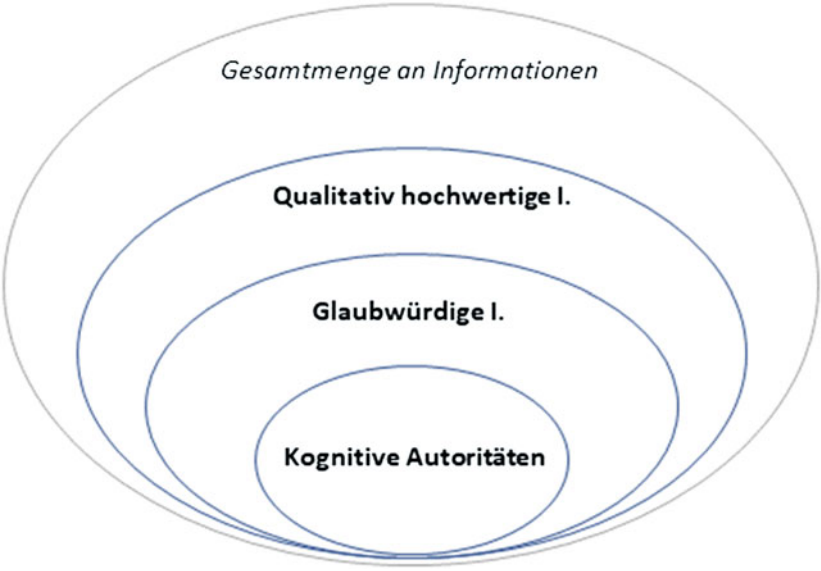

Abbildung 2.1 Unterscheidung zwischen wahrgenommener Qualität, Glaubwürdigkeit und kognitiver Autorität von Informationen

Allgemein können die Konzepte Glaubwürdigkeit und Autorität als eng miteinander verknüpft verstanden werden. Dennoch stellt sich die Frage, ob Glaubwürdigkeit als Basis für kognitive Autorität dient oder ob kognitive Autorität das grundlegende Konzept hinter Glaubwürdigkeit darstellt, wie Rieh & Belkin (1998) schreiben.

Rieh & Danielson (2007) stellen in ihrem ausführlichen Literaturüberblick die Zusammenhänge von Glaubwürdigkeit und verwandten Konzepten wie Qualität, Autorität, Vertrauen und Persuasion dar und zeigen auf, dass Glaubwürdigkeit als multidisziplinäres Framework zu verstehen ist, denn nicht nur in den informationswissenschaftlichen Bereichen Informationssuche und IR dient dieses Konzept als Forschungsgegenstand, sondern beispielsweise auch in den Medienwissenschaften, Gesundheitswissenschaften und der Kaufverhaltensforschung. Die Autoren erläutern, dass Informationen, die als glaubwürdig beurteilt werden, eine Teilmenge von Informationen sind, die als qualitativ hochwertig wahrgenommen werden und verorten Informationen, von denen informationssuchende Personen kognitive Autoritäten ableiten können, als Teilmenge von

glaubwürdigen Informationen (Rieh & Danielson, 2007, S. 345). Abbildung 2.1 veranschaulicht diese Einteilung von Informationen in einzelne Teilmengen.

Die Arbeit von Olaisen (1990) betraf zwar nicht den Kontext der Websuche, ist aber vermutlich die erste empirische Studie, die explizit auf kognitive Autorität und Glaubwürdigkeit von elektronischen Informationen bei der Informationssuche abzielt (Rieh & Danielson, 2007, S. 317). Olaisen (1990) untersuchte die Faktoren zur Beurteilung von Informationsqualität von Personen im Finanz- bzw. Versicherungssektor Norwegens mithilfe von Fragebögen und Interviews. Das Ergebnis besteht unter anderem aus einem Ranking der fünf Faktoren, die einer Quelle kognitive Autorität zuschreiben: Glaubwürdigkeit im Sinne von Vertrauenswürdigkeit, Relevanz, Zuverlässigkeit, Validität und Bedeutung über die Zeit hinaus (Olaisen, 1990, S. 119). Zu beachten ist hierbei die Nennung von Relevanz als Faktor auf einer Ebene mit Vertrauenswürdigkeit und Zuverlässigkeit, was den Erkenntnissen der informationswissenschaftlichen Forschung zu Relevanz als Konzept und Relevanzkriterien widerspricht – werden doch Vertrauenswürdigkeit, Zuverlässigkeit und Validität als Kriterien für die Bewertung von Relevanz erachtet (vgl. Abschnitt 2.1.1).

Rieh (2002) untersuchte, wie Menschen Qualität und kognitive Autorität bei der Websuche evaluieren. Sie argumentiert, dass Menschen die Qualität und Autorität von gedruckten Materialien im Gegensatz zu Dokumenten im Web generell leichter beurteilen könnten, weil ihnen Wissen aus langjährigen Erfahrungen über traditionelle Informationsquellen zur Verfügung stünden, um die ihnen bekannten Indikatoren für Qualität (z. B. redaktionelle Auswahl) und Autorität (z. B. Autoren, Verlage) heranziehen zu können. Für Informationen im Web würden diese Indikatoren nicht in demselben Maß gelten, da die Inhalte vor Veröffentlichung nicht zwingend eine Qualitätskontrolle (im Gegensatz zu wissenschaftlichen Artikeln in seriösen Fachzeitschriften) durchlaufen. Auch Daten über die Verantwortlichen des Inhalts (z. B. im Impressum eines Webauftritts) könnten fehlerhaft sein. Zur Überprüfung bzw. Bewertung von Qualität und Autorität im Web müssten informationssuchende Personen einen höheren Aufwand betreiben als in anderen Information Retrieval-Systemen (Rieh & Belkin, 1998).

Rieh folgte R. S. Taylor (1986) in der Definition von Informationsqualität, nach der Qualität ein Nutzerkriterium ist und bestimmte Werte (*values*) beinhaltet: Richtigkeit (*accuracy*), Vollständigkeit (*comprehensiveness*), Aktualität (*currency*), Zuverlässigkeit (*reliability*) und Gültigkeit (*validity*). Diese Werte sind größtenteils schwer bestimmbar, denn „we tend to be suspicious of a system or a package which needs to advertise its reliability or its accuracy by words only. These are characteristics earned over time and by reputation“ (R. S. Taylor, 1986, S. 62). Für ihre Studie operationalisierte Rieh (2002) allerdings das Konzept von

Qualität unter Hinzunahme von Nützlichkeit[10] und verwendet damit – ähnlich wie Olaisen (1990) den Begriff Relevanz – ein Konzept, das sich von den anderen, als Kriterien der Relevanzbewertung, in seiner Bedeutung unterscheidet:

> At an operational level, information quality is identified as the extent to which users think that *the information is useful, good, current, and accurate*. Cognitive authority is operationalized as to the extent to which users think that *they can trust the information*. (S. 146)

Rieh untersuchte den interaktiven IR-Prozess im Web mithilfe von menschlichen Jurorinnen und Juroren, die auf Basis ihrer eigenen Suchanfragen zu vier verschiedenen Aufgaben Webseiten evaluierten. Die Aufgabenbeschreibungen enthielten Formulierungen über die gewünschten relevanten Suchergebnisse entsprechend der operationalisierten Definition von Qualität und Autorität, wie „good papers", „useful information", „credible information", „best price" (Rieh, 2002, S. 149):

> (1) For the research project in which you are currently engaged, you would like to find some good papers which are new to you, which you think will be useful (*research* task).
> (2) You are planning for the next conference that you are going to attend, and would like to find useful information about hotels, restaurants, and features of interest in that city (*travel* task).
> (3) A friend of yours has just been diagnosed as having schistosomiasis, and you want to find credible information about the disease itself, and the best methods of treatment (*medicine* task).
> (4) You've decided that you want to buy a new computer to use at home, and now you need to find the best price for it (*computer* task).

Obwohl die Begriffe Relevanz oder relevant in den Aufgabenstellungen nicht auftauchen, ist das Gesamtkonzept von Relevanz impliziert durch Kriterien wie Güte und Vertrauenswürdigkeit. Nützlichkeit (*usefulness*) wird für die Aufgabe (2) verwendet, welche konkrete faktenorientierte Antworten außerhalb eines akademischen Kontexts verlangt. Für die Aufgabe (1) im akademischen Kontext hingegen wird Güte als Kriterium genannt, während die Aufgabe (3) auf vertrauenswürdige Informationen im Gesundheitsbereich abzielt und Aufgabe (4) als navigations- bzw. transaktionsorientierte Suche verstanden werden kann, weil

[10] Auf das Konzept der Nützlichkeit (*usefulness*) als eine Form von Relevanz bzw. alternatives Maß zur Bewertung der Retrieval-Effektivität wird in Abschnitt 3.1 näher eingegangen.

nur der beste Preis gefunden werden soll (im Gegensatz zu beispielsweise den besten drei Preisen).

Die Daten wurden mittels qualitativer Befragungsverfahren (Think-Aloud-Protokollen, Anschlussinterviews) und einer technikgestützten Beobachtung (Logfiles der Suchsitzungen) erhoben. An der Studie nahmen 16 Fakultätsangehörige bzw. Promovierende teil. Es wurden die Daten von 15 Teilnehmenden ausgewertet, die zu insgesamt 1.321 evaluierten Webseiten vorlagen.

Die Erkenntnisse ihrer Studie fasst Rieh (2002) in einem Modell zur Bewertung der Informationsqualität und kognitiven Autorität (*Model of Judgment of Information Quality and Cognitive Authority*) im Kontext der Websuche zusammen (Abbildung 2.2). In dem Modell wird der Bewertungsprozess als zentraler Aspekt im gesamten interaktiven Prozess der Informationssuche im Web hervorgehoben, wobei dieser aus mehreren Iterationen von *predictive judgments* (Bewertung von Suchergebnissen) und *evaluative judgments* (Bewertung der Webseiten) bestehen kann.

Das Modell stellt den Zusammenhang zwischen den Kriterien Informationsqualität und kognitiver Autorität und deren vielfältige Facetten im Kontext verschiedener Faktoren wie Aufgabe, Nutzerwissen sowie den Attributen (Charakteristika) der zu bewertenden Informationsobjekte und Quellen her. Positiv hervorzuheben ist, dass dabei zwischen den Zeitpunkten der Bewertung bzw. der damit einhergehenden Bewertungsgrundlage (Suchergebnis oder Webseite) explizit unterschieden wird, wobei ein direkter Zusammenhang zwischen der Bewertungsgrundlage und den Kriterien hergestellt wird. So unterscheidet Rieh die Eigenschaften von Informationen im Web anhand zweier Kategorien: (a) Eigenschaften des Informationsobjekts (Charakteristika von Informationsobjekten) sind der Typ, Titel, Inhalt, die Organisation/Struktur, Präsentation, Abbildungen und Funktionalität; (b) zu den Eigenschaften der Quelle (Charakteristika von Quellen) zählen die URL-Domäne, der Typ, die Reputation der Quelle, Referenzen zu Autor/Urheber und ob es sich um die Arbeit einer einzelnen Person oder mehrere Personen als Kollektiv handelt. Zugleich wird eine klare Trennung bei der Zuweisung zu den Kriterien vorgenommen: Anhand der Eigenschaften des Informationsobjekts beurteilen informationssuchende Personen die Informationsqualität, die Eigenschaften der Quelle dienen als Basis für die Bewertung der kognitiven Autorität.

Allerdings zeigen die Studienergebnisse, dass die Bewertung der Informationsqualität und kognitiven Autorität stark aufgabenabhängig war. Informationsqualität scheint von größerer Bedeutung bei den Aufgaben zu dem Forschungsprojekt und dem Computerkauf gewesen zu sein, kognitive Autorität hingegen bei

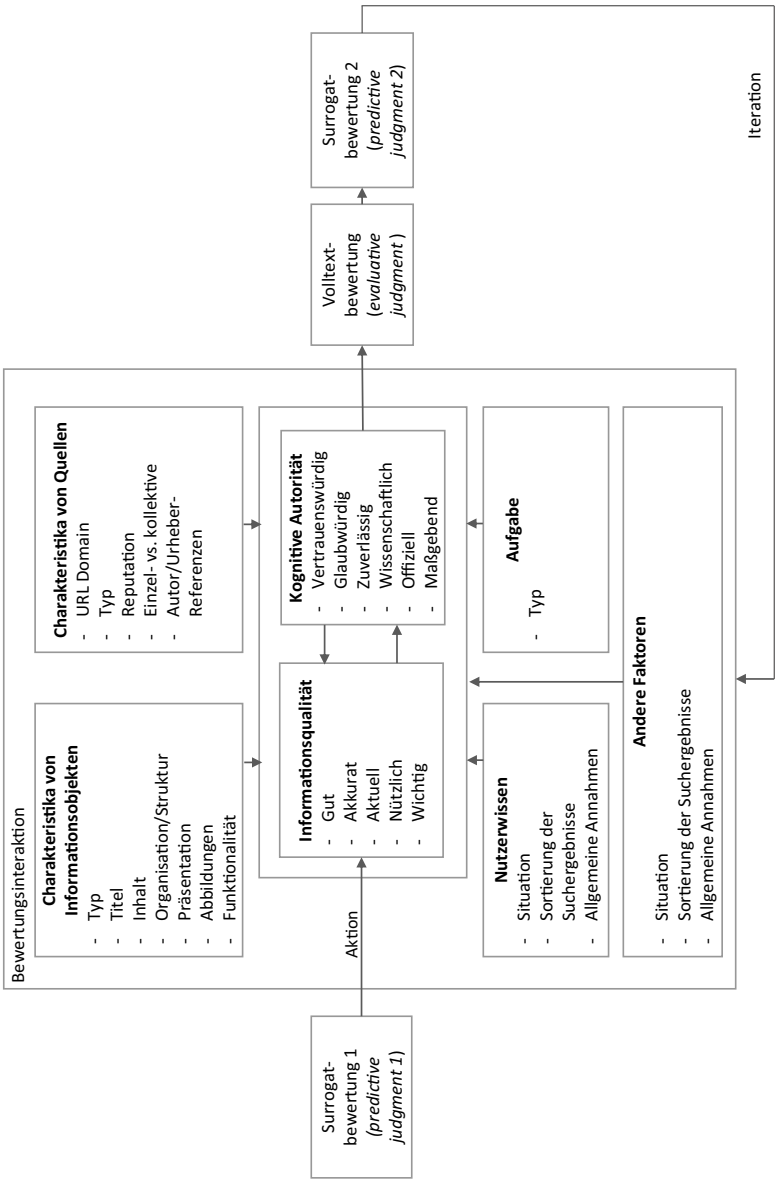

Abbildung 2.2 Modell zur Bewertung der Informationsqualität und kognitiven Autorität (übersetzt aus Rieh, 2002, S. 158)

der medizinischen Aufgabe (Rieh, 2002, S. 151). Aufgrund der relativ kleinen Stichprobe (n = 15) wären jedoch weitere Untersuchungen notwendig, um aussagekräftige Ergebnisse zu erhalten.

Riehs Modell zeigt auf, anhand welcher Hinweise bzw. Merkmale eines Suchergebnisses während der Websuche dessen Qualität und kognitive Autorität abgeleitet werden. Zu unterscheiden sind dabei die Erwartungen an die eigentliche Webseite, die sich in Form der *predictive judgments* ausdrücken, und die tatsächliche Evaluierung der aufgerufenen Webseite, im Rahmen derer überprüft wird, inwieweit die Erwartungen erfüllt wurden, d. h. das *predictive judgment* dem *evaluative judgment* entspricht. Rieh zieht folgende Schlussfolgerung aus den Ergebnissen ihrer Studie:

> Web users would make their predictive judgments more effectively if they could see more clues that indicate the facets of information quality and cognitive authority. […] If information objects and sources on results page were more detailed, users would make better predictive judgments, and they would be less likely to have to return to the search results to open another page. This study confirms this, showing that information about sources at institutional (name or type of source) and individual (author/creator) levels could be very helpful for users who tend to make predictive judgments based on characteristics of sources. (2002, S. 159)

Erwartungen an ein Dokument auf Basis seines Surrogates hinsichtlich dessen Qualität werden bereits bei Wang & Soergel (1998) thematisiert. Sie weisen das Kriterium der erwarteten Qualität (*expected quality*) als geschätzte Güte eines Dokuments im Zusammenhang mit der Qualität eines Journals und einer Autorin oder eines Autors nach. Die Nutzung dieses Kriteriums ist zwangsläufig auf die Surrogatbewertung (*predictive judgment*) beschränkt, sie erfolgt „before consulting the full document“ und setzt dabei auf die Bewertung der thematischen Relevanz auf – „if the topic did not match, quality was not judged“ (Wang & Soergel, 1998, S. 123).

Hinweise für die Bewertung der kognitiven Autorität eines Textes sieht Olaisen (1990) in der kognitiven Autorität seiner Autorin oder seines Autors und in der Reputation des Verlags, die ebenfalls Elemente eines Surrogates darstellen und entsprechende Kenntnisse über die Autorin, den Autor und den Verlag voraussetzen.

Wenn eine informationssuchende Person jedoch nicht über die zur Beurteilung von Qualität, Glaubwürdigkeit und Autorität notwendigen Kenntnisse verfügt, stellt sich die Frage, anhand welcher Elemente sie die Glaubwürdigkeit von Informationsobjekten – bewusst oder unbewusst – ableitet. An dieser Stelle kommt das Konzept der Popularität zum Tragen: So simulieren Anbieter von Suchsystemen

über Popularität die Glaubwürdigkeit ihrer Suchergebnisse, indem sie Popularitätsfaktoren in ihre Rankingalgorithmen integrieren (Lewandowski, 2012). Im nachfolgenden Abschnitt wird das Konzept von Popularität als Indikator für die (erwartete) Qualität eines Suchergebnisses näher betrachtet. Aufgrund fehlender Studien zum Einfluss von Popularität auf die Relevanzbewertung, können zur Betrachtung entsprechende Erkenntnisse empirischer Studien nicht herangezogen werden.

2.1.2.2 Popularität als Indikator für Qualität

Das Konzept der Popularität im Kontext der Informationssuche lässt sich mit dem Konzept der *Weisheit der Vielen (wisdom of crowds)*[11] beschreiben: Das Wissen und die Erfahrungen von vielen können als bedeutsamer erachtet werden als das Wissen des Einzelnen; das bedeutet, je größer die Anzahl an Personen, die ein Dokument als relevant erachten, desto höher ist die Wahrscheinlichkeit, dass dieses Dokument für einen weiteren Einzelnen ebenfalls relevant ist. Obwohl diese Darlegung weder den hohen Grad an Subjektivität (Beziehung zum individuellen Informationsbedürfnis) noch die Kontextabhängigkeit von Relevanz aus nutzerbasierter Sicht berücksichtigt, kann das Konzept von Popularität in der Websuche als äußerst erfolgreich beurteilt werden: Seit der Einführung des PageRank-Verfahrens (Page et al., 1998) stellt es eines der grundlegenden Konzepte hinter dem Ranking bekannter Websuchmaschinen wie Google dar.

Im Zusammenhang mit dem Ranking in (wissenschaftlichen) Bibliothekskatalogen taucht das Konzept der Popularitätsfaktoren bereits bei Lewandowski (2009) auf und wird von Behnert & Lewandowski (2015) – inspiriert von Rankingfaktoren bei der Websuche – weiter vertieft, die beispielsweise Klickhäufigkeiten, Verweildauer, Nutzungshäufigkeiten und Zitationszahlen (im Gegensatz zu Impact-Kennzahlen wie der h-Index) als potenzielle Faktoren nennen. Diese

[11] Surowiecki (2005) erläutert anhand des Konzepts der Weisheit der Vielen laut Untertitel seines Buches, „[w]arum Gruppen klüger sind als Einzelne und wie wir das kollektive Wissen für unser wirtschaftliches, soziales und politisches Handeln nutzen können“. Die Entdeckung der *Wisdom of Crowds* geht zurück auf Galton (1907), der in seinen Beitrag „Vox populi“, beschreibt, wie bei seiner Untersuchung zu einer Vielzahl von einzelnen Schätzungen über das Gewicht eines Ochsen die Weisheit der Vielen nur um einen Prozentpunkt von dem korrekten Ergebnis abwich.

Kennzahlen können als Popularitätsdaten[12] (*popularity data*) bezeichnet werden, wie es Richardson et al. (2010) im Zusammenhang mit dem Ranking von Suchergebnissen erstmals in ihrem 2005 angemeldeten US-Patent taten:

> The Subject application relates to a system(s) and/or methodology that facilitate using popularity data to improve the ranking of objects and ultimately, to obtain more relevant search results. More specifically, the system and method involve tracking which objects have been viewed, visited, or accessed to determine a measure for each and using the measure or some function thereof to determine a popularity based ranking for each of the objects. (Richardson et al., 2010, Sp. 1)

Zudem wird deutlich, dass das Ranking anhand dieser Popularitätsdaten letztendlich auf die Verbesserung der Qualität abzielt, denn es ist das Ziel, „to improve or enhance the quality and/or accuracy of search results" (Richardson et al., 2010, Sp. 2). Mithilfe solcher Popularitätsdaten wird die Reihenfolge, in der die Suchergebnisse präsentiert werden, beeinflusst und damit auch die Relevanzbewertung, ohne diese Daten der informationssuchenden Person explizit anzuzeigen. So konnten etliche Studien zeigen, dass Menschen die Suchergebnisse auf den obersten Trefferpositionen bevorzugen, weil sie dem Ranking von Suchmaschinen im Web viel Vertrauen entgegenbringen und sich zumeist darauf verlassen, dass die Ergebnisse mit der für sie höchsten Relevanz auf den ersten Positionen angezeigt werden (vgl. z. B. C. Barry & Lardner, 2011; Jansen & Spink, 2006; Pan et al., 2007; Schultheiß et al., 2018).

Als expliziter Bestandteil der Suchergebnisdarstellung sind Popularitätsdaten aus den sozialen Medien (*Likes*) und aus dem E-Commerce-Bereich bekannt, zum Beispiel die Sterne-Bewertungen (in Relation zur Anzahl der Produktrezensionen) bei Amazon.de und ähnlichen Online-Shops.[13]

Die zum Zeitpunkt der Erfassung des Forschungsstands einzige bekannte Studie zu Relevanzkriterien, in der explizite Popularitätsdaten berücksichtigt und Popularität als Relevanzkriterium erachtet werden, ist die Tagebuchstudie von

[12] Der Begriff Popularitätsdaten taucht bereits bei Lewandowski (2010) im Zusammenhang mit Bibliothekskatalogen auf, bezeichnet dort allerdings eine andere Art von Daten, die anfrageunabhängig sind, wie beispielsweise die gruppierten Titel eines ausgewählten (angesehenen) Verlages.

[13] In ähnlicher Weise funktionieren Bestsellerlisten für den konventionellen Buchhandel, die basierend auf den Verkaufszahlen in bestimmten Bereichen erstellt und als Indikator für Qualität mittels Popularität erachtet werden.

Albassam & Ruthven (2018). Diese Studie bezieht sich jedoch nicht auf Suchergebnisse eines textbasierten IR-Systems[14]: Im Zusammenhang mit der Auswahl von YouTube-Videos im Freizeitkontext operationalisieren die Autoren Popularität als die Anzahl von Aufrufen und Gefällt-mir-Anzeigen (*Likes*) und fanden heraus, dass das Kriterium[15] Popularität als Beurteilung der (vermuteten) Qualität genutzt wurde: „It could be noticed from some responses that participants predict some level of video's quality based on its popularity. For example, 'it had over 2 million views so I could safely assume it was a reliable link'…" (Albassam & Ruthven, 2018, S. 72). Des Weiteren zeigte sich, dass viele der allgemeinen Kriterien, die aus früheren Studien zu Relevanzkriterien bekannt sind, ebenfalls unter den Kriterien zur Auswahl der Videos sind; dagegen konnten keine Hinweise auf Kriterien, welche die Autoren eher dem akademischen Kontext zuordnen wie beispielsweise Autorität, gefunden werden. Vor diesem Hintergrund stellt sich die Frage, inwiefern sich die Kriterien Glaubwürdigkeit, kognitive Autorität und Popularität generell auf die Suche nach und Bewertung von Surrogaten in akademischen Suchsystemen übertragen lassen.

2.1.3 Kriterien im Kontext akademischer Suchsysteme

Die in Abschnitt 2.1.1 vorgestellten, allgemein geltenden Kriterien sind auf den Kontext[16] der Informationssuche in akademischen Suchsystemen übertragbar; auch hier bildet die Beurteilung der thematischen Relevanz die Basis der Relevanzbewertung. Die Bewertung von Suchergebnissen in akademischen Suchsystemen unterscheidet sich von der Bewertung während der Websuche, d. h. in allgemeinen Websuchmaschinen, dahingehend, dass die Suchergebnisse andere

[14] Die Studie zu Relevanzbewertungen von Videos auf YouTube von Albassam & Ruthven (2018) wird im nachfolgenden Abschnitt 2.2 nicht näher betrachtet und war nicht Bestandteil der Analyse der ausgewählten Studien zu Relevanzkriterien.

[15] Die Teilnehmenden waren aufgefordert, unter anderem die Gründe (*reasons*) für die Auswahl bzw. das Ansehen der Videos in das halbstrukturierte Tagebuch zu schreiben. Hier zeigt sich erneut das Problem der fehlenden definitorischen Abgrenzung, denn die Antworten können ein sehr breites Spektrum an Einflussparametern darstellen, das einen ebenso breiten Interpretationsspielraum im Zuge des Codierens und der inhaltlichen Analyse zulässt.

[16] Der Begriff Kontext im Zusammenhang mit der Informationssuche bezeichnet im Rahmen dieser Arbeit die Art des Suchsystems bzw. die Art der gesuchten Informationen. Diese Unterscheidung wird insbesondere in Abschnitt 2.2 bei der Einordnung ausgewählter Studien deutlich, bei der neben dem akademischen Kontext (Suche nach wissenschaftlichen Informationen bzw. Suche in akademischen Suchsystemen) beispielsweise auch zwischen dem schulischen und beruflichen Kontext differenziert wird.

Informationen über das eigentliche Dokument in einer anders strukturierten Darstellungsform bieten, die den Bedürfnissen und Erwartungen der Zielgruppe (Personen mit einem Informationsbedürfnis im akademischen Kontext) eher entsprechen.

Es ist davon auszugehen, dass das Kriterium der Qualität einen besonderen Stellenwert im akademischen Kontext einnimmt. So wählten beispielsweise Rieh & Belkin (1998) sowie Rieh (2002) für ihre Untersuchungen zur Bewertung der Informationsqualität und kognitiven Autorität gezielt Wissenschaftlerinnen und Wissenschaftler als Grundgesamtheit aus unter der Annahme, dass diese sich mit größerer Wahrscheinlichkeit stärker als andere Bevölkerungsgruppen mit Informationsqualität und kognitiver Autorität auseinandersetzen dürften.

Qualität ist ein unentbehrliches Konzept in der Wissenschaft, das innerhalb und außerhalb des Wissenschaftssystems beachtet wird. Innerhalb der eigenen wissenschaftlichen Community bewerten Kolleginnen und Kollegen die Qualität von Publikationen und Forschungsanträgen im Rahmen des Peer Review, für die eigene Arbeit bewerten Forschende die Qualität von Publikationen anderer; außerhalb des Wissenschaftssystem beurteilen verschiedene Akteure aus Politik, Medien und Gesellschaft die Qualität wissenschaftlicher Forschung (Döring & Bortz, 2016, S. 84).

Für die Beurteilung wissenschaftlicher Güte gelten vier Qualitätskriterien, die Döring & Bortz (2016) im Zusammenhang mit Qualität in der empirischen Sozialforschung erläutern. Allerdings sind sie als „paradigmen- und disziplinübergreifend zu betrachten, müssen jedoch disziplin- und paradigmenspezifisch konkretisiert werden“ (S. 89–90): (a) inhaltliche Relevanz des Forschungsproblems in Hinblick auf die theoretische/wissenschaftliche Relevanz im einem bestimmten Forschungsfeld bzw. praktische Relevanz in der Anwendungsforschung, (b) methodische Strenge im wissenschaftlichen Forschungsprozess, (c) ethische Strenge hinsichtlich wissenschafts- und forschungsethischer Aspekte, (d) Präsentationsqualität in Hinblick auf die Dokumentation des Forschungsprozesses und Darstellung seiner Ergebnisse (Döring & Bortz, 2016, S. 89 ff.).

Vor diesem Hintergrund stellt sich die Frage, ob und in welcher Weise die Kriterien für Qualität in der Wissenschaft als Relevanzkriterien bei der Bewertung von Suchergebnissen in akademischen Suchsystemen herangezogen werden können. Hjørland & Christensen, (2002) sehen einen direkten Zusammenhang zwischen den Vorgaben innerhalb und außerhalb des Wissenschaftssystems und dem Kontext sowie der Situation einer informationssuchenden Person (*situational relevance)*:

> [T]he basic paradigms have been developed by the research institutions and universities where professionals are trained. Such institutions may be more or less depending on and indirectly influenced by financial support from outside sources. Those paradigms set the frames within which the situational relevance may be defined. They also influence the terminology, the research methodology, relevance criteria, the citation patterns, the publication – and the retrieval system. The more influential the view, the more dominating will its conceptualization and relevance criteria be. The dominating view looks "natural," and minority conceptions tend to look strange and less professional. (Hjørland & Christensen, 2002, S. 962)

Studien zu Relevanzkriterien im akademischen Kontext bauen oft auf den Arbeiten von Barry und Schamber auf, deren Bedeutung bereits im Zusammenhang mit allgemeinen Kriterien betont wurde (vgl. Abschnitt 2.1.1). Barry (1994) führte Interviews mit 18 Studierenden bzw. Angehörigen des Lehrpersonals einer Universität aus fünf verschiedenen Fachbereichen durch: Geographie und Anthropologie, Psychologie, Englisch, Geschichte und Literatur. Sie ließ die Teilnehmenden Suchergebnisse in DIALOG[17] beurteilen, indem diese die Teile der Surrogate zu den Dokumenten, die sie näher betrachten wollten, markierten; im Anschluss wurden sie von der Forschungsleitung nach den Gründen für die Auswahl der markierten Elemente bzw. Bereiche befragt. Als Ergebnis der Datenauswertung gruppierte Barry 23 Kriterienkategorien und ordnete sie 7 Kriterienklassen zu (Tabelle 2.3). Unter ihnen finden sich neben dem Kriterium für Qualität (*Source quality*) auch Hinweise auf Kriterien der Autorität bzw. kognitiven Autorität (*Source reputation/visibility*, *Relationship with author*).

Tabelle 2.3 Kriterien aus der Studie von Barry (1994, S. 154)

Grouping	Criterion categories
Criteria pertaining to the information content of documents	Depth/scope Objective accuracy/validity Tangibility Effectiveness Clarity Recency

(Fortsetzung)

[17] Der Datenbankanbieter DIALOG mit einem breiten inhaltlichen Spektrum an Datenbanken gehörte in der Vergangenheit zu Thomson Scientific und ist inzwischen Teil von ProQuest. Weitere Informationen sind zu finden unter www.dialog.com.

Tabelle 2.3 (Fortsetzung)

Grouping	Criterion categories
Criteria pertaining to the user's previous experience and background	Background/experience Ability to understand Content novelty Source novelty Stimulus document novelty
Criteria pertaining to the user's beliefs and preferences	Subjective accuracy/validity Affectiveness
Criteria pertaining to other information and sources within the information environment	Consensus External verification Availability within the environment Personal availability
Criteria pertaining to the sources of documents	Source quality Source reputation/visibility
Criteria pertaining to the document as a physical entity	Obtainability Cost
Criteria pertaining to the user's situation	Time constraints Relationship with author

Choi & Rasmussen (2002) verwendeten diese Kriterien als Vorlage zur Ermittlung einer Kriteriengewichtung, Tang & Solomon (2001) orientierten sich an den identifizierten Kriteriengruppen. Beresi et al. (2010) ließen für ihre Studie Versuchspersonen aus drei unterschiedlichen Fachdisziplinen/-bereichen (Informatik, Informationsmanagement, Pharmazie) Surrogate bewerten, um unter anderem mögliche Unterschiede zwischen den Personen der verschiedenen Fachdisziplinen bezüglich der Verwendung von Relevanzkriterien aufzudecken. Sie nutzten die von Barry (1994) und Barry & Schamber (1998) ermittelten Relevanzkriterien nach und stellten fest:

> We can immediately observe that tangibility and depth/scope/specificity are the most mentioned criteria […] while participants from the School of Computing have a distinguishable preference for tangible data, members of the other two schools prefer other aspects of the information such as its depth, scope and specificity. Furthermore, we can also observe that members from all three schools share the same interest (in terms of proportions) for the novelty of the documents found. (Beresi et al., 2010, S. 202–203)

Zudem verweisen sie auf die Notwendigkeit der Kriterien als messbare Variablen und zeigen beispielhaft Möglichkeiten der Operationalisierung für die zwei häufigsten genannten Kriterien:

> Relevance criteria are not theoretical concepts, but rather tangible and operationalising them can potentially impact positively on search services. […] If, and only if, we can measure them. *Tangibility*, may be approximated, for instance, by looking at the number of tables in a document, and *depth/scope/specificity*, by looking at the number of pages in a document (document length has been mentioned frequently as a relevance criteria [*sic*]). (Beresi et al., 2010, S. 206, Kursivdruck im Original)

Die oben beschriebenen vier Kriterien der wissenschaftlichen Qualität lassen sich in Gänze nur anhand des tatsächlichen Dokumenteninhalts ableiten, ein Surrogat hingegen kann zur Bewertung der *erwarteten* Qualität (vgl. Abschnitt 2.1.2.1) dienen. Stellvertretend für den Dokumenteninhalt beinhalten Surrogate in akademischen Suchsystemen heutzutage zusätzliche Elemente, anhand derer Relevanzkriterien bzw. Kriterien für die erwartete Qualität abgeleitet werden können. Diese zusätzlichen Elemente sind beispielsweise Popularitätsdaten (vgl. Abschnitt 2.1.2.2). Im akademischen Kontext können die Anzahl von Zitationen eines Werkes, einer Autorin oder eines Autors, die als Faktoren für die Qualität und wiederum die Glaubwürdigkeit gesehen werden können, als Popularitätsdaten dienen. Downloadhäufigkeiten oder Ausleihzahlen – wie speziell im Bibliothekskontext verankert – kommen als Indikatoren der Nutzungsintensität ebenfalls als Popularitätsdaten im akademischen Kontext infrage, auch wenn sie nicht explizit als Bestandteil der Suchergebnispräsentation angezeigt, sondern als Rankingfaktor in bibliothekarischen Informationssystemen herangezogen werden (Plassmeier et al., 2015).

Der Stellenwert solcher Popularitätsdaten, wie die Anzahl von Zitationen, ist jedoch vor dem Hintergrund des Matthäuseffekts[18] kritisch zu hinterfragen: Werke anerkannter Persönlichkeiten erlangen eine höhere Anerkennung bereits dadurch, dass ihre Autorinnen und Autoren eine gewisse Reputation besitzen. Die explizite Anzeige solcher zusätzlichen Informationen in Suchergebnissen akademischer Suchsysteme bewirkt möglicherweise, dass bereits viel zitierte Werke als qualitativ wertvoller beurteilt werden, obwohl die Beurteilung der

[18] Der Matthäuseffekt in der Wissenschaft besteht darin, dass Wissenschaftlerinnen und Wissenschaftlern von beachtlichem Ansehen für wissenschaftliche Beiträge mehr Anerkennung wiederfährt und dass Wissenschaftlerinnen und Wissenschaftlern, die sich noch nicht profiliert haben, diese Anerkennung vorenthalten wird (Merton, 1968). Die Bezeichnung geht zurück auf ein Gleichnis des Matthäusevangeliums in der Bibel und dessen heutzutage sinngemäß wahrgenommener Bedeutung: „Wer hat, dem wird gegeben."

tatsächlichen Qualität in Hinblick auf die oben genannten Kriterien der wissenschaftlichen Qualität auf der Basis des Dokumenteninhalts unter Umständen von der der *erwarteten* Qualität stark abweicht. Im besten Fall entspricht die erwartete Qualität voll und ganz der tatsächlichen Qualität, im schlechtesten Fall stimmt sie überhaupt nicht überein. Dass Diskrepanzen bei der Relevanzbewertung von Surrogaten und der Relevanzbewertung der dazugehörigen Volltexte durchaus vorkommen, zeigt die Studie von Lewandowski (2008) zur Retrieval-Effektivität von Websuchmaschinen unter Berücksichtigung von Snippets[19], die als Surrogate der verlinkten Webseiten und anderer Dokumente dienen. Da Popularitätsdaten auf die Anzeige in Surrogaten begrenzt sind, bedeutet dies, dass Popularität als Relevanzkriterium im Zusammenhang mit *predictive judgments* zum Tragen kommt; jedoch ist nicht auszuschließen, dass Popularität auch das *evaluative judgment* beeinflusst.

Im Zusammenhang mit Qualität und kognitiver Autorität wurde zuvor bereits die Erkenntnis von Rieh (2002) zitiert, dass weitere, detailliertere Informationen, die einen Hinweis auf die zu erwartende Qualität eines Suchergebnisses liefern können, zu einem besseren *predictive judgment* führen (vgl. Abschnitt 2.1.2.1). Vor diesem Hintergrund werden im nachfolgenden Abschnitt Surrogate als Bewertungsgrundlage in Studien zu Relevanzkriterien (im akademischen Kontext) in den Fokus genommen.

2.1.4 Surrogate als Grundlage der Bewertung

In bibliothekarischen Informationssystemen werden Dokumente seit jeher anhand ihrer Metadaten zu den gedruckten Materialien repräsentiert. Diese Dokumentrepräsentationen (Surrogate) liefern somit allein durch die bibliographischen Angaben wichtige Merkmale für die erste Relevanzbewertung, die von einer informationssuchenden Person vorgenommen wird. Die Repräsentation von Dokumenten als Bewertungsgrundlage stellt einen wichtigen Aspekt in Studien zu Relevanzkriterien dar. In einigen von ihnen wurde gezielt die Verwendung von Relevanzkriterien zu unterschiedlichen Zeitpunkten, also vor und nach dem Zugang zum Volltext, untersucht (z. B. Crystal & Greenberg, 2006; Tang & Solomon, 2001). In Hinblick auf die Relevanzbewertung kann dahingehend zwischen *predictive* und *evaluative judgments* unterschieden werden – eine Einteilung von Bewertungen, die auf die entscheidungspsychologischen Arbeiten von Hogarth

[19] Snippets bezeichnen kurze Beschreibungstexte der Suchergebnisse in der von der Websuchmaschine präsentierten Ergebnisliste.

(1987) zurückgeht und bei Rieh (2002) Anwendung im Kontext der Websuche findet (vgl. Abschnitt 2.1.2.1).

In Bezug auf den Zeitpunkt der Bewertung erfolgt die erste Bewertung anhand des Surrogats, noch bevor die suchende Person entschieden hat, welche Auswahl sie hinsichtlich des Volltextaufrufs treffen wird:

> In the model […], a predictive judgment guides a decision about what kind of action the user is going to take given multiple choices (alternatives). As a result of this judgment, a new Web page is presented to the user, and when she/he looks at it, an evaluative judgment is made. (Rieh, 2002, S. 146–147)

Mit ihrer Studie zeigte Rieh (2002), dass ein (positives) *predictive judgment* zu dem Aufrufen einer Webseite führt, anhand derer das *evaluative judgment* vorgenommen wird. Während der Bewertungen nannten die Teilnehmenden unter Anwendung der Methode des lauten Denkens verschiedene Schlüsselwörter, welche die Erwartungshaltung des *predictive judgment* bzw. die Gegebenheit des *evaluative judgments* ausdrücken:

> The keywords and phrases that appeared often in the subjects' predictive judgments included: "It *would be* a good search engine;" "It *is likely to be* good;" "It *will* give me reliable databases;" "It *sounds like* a generic name." Note that the phrases indicate expectations, anticipations, and predictions regarding the page that the subjects decided to look at. (Rieh, 2002, S. 150; Kursivdruck im Original enthalten)
> On the other hand, the keywords and phrases which appeared in the evaluative judgments included: "It *turned out* it wasn't what I expected;" "I *did* find this article interesting;" "It *looks* scholarly;" "It *seems* to be a kind of authentic organization." Here, the phrases indicate evaluations of the page based on the information presented within. (Rieh, 2002, S. 151; Kursivdruck im Original enthalten)

Diese Notwendigkeit zur Unterscheidung zwischen *predictive* und *evaluative judgments* sieht auch Taraborelli, (2008), der kritisiert, dass Studien zu Glaubwürdigkeit im Web bis dato die Rolle von *predictive judgments* vernachlässigen, obwohl eine Fülle an zentralen Hinweisen (*proximal cues*) im Web bestehen, anhand derer Informationsquellen ausgewählt werden. Solche *proximal cues* gelten als wesentliche Bestandteile für *information scent*[20] (Pirolli & Card, 1999), einer Kennzahl der wahrgenommenen Profitabilität einer externen Quelle vor ihrer Auswahl. Auch Rieh & Danielson (2007) betonen, dass Beurteilungen von

[20] Pirolli & Card (1999) definieren: „Information scent is the (imperfect) perception of the value, cost, or access path of information sources obtained from proximal cues, such as bibliographic citations, WWW links, or icons representing the sources" (S. 646).

Glaubwürdigkeit an menschliche Bewertungen gebunden sind und Attribute von Dokumenten lediglich Hinweise für solche Bewertungen liefern können. Da *predictive judgments* unabhängig davon, ob ein Dokument nach dessen Relevanz, Nützlichkeit, Glaubwürdigkeit oder Autorität zu beurteilen ist, auf den Elementen eines Surrogates beruhen, können sie lediglich Vermutungen oder Schätzungen über die tatsächliche – wenngleich immer noch subjektiv bewertete – Nützlichkeit, Relevanz oder Qualität darstellen, die erst mit den *evaluative judgments* der Dokumenteninhalte zu einem späteren Zeitpunkt verifiziert werden. Auch in traditionellen IR-Studien zur Evaluierung der Retrieval-Effektivität von Suchsystemen blieben *predictive judgments* lange unberücksichtigt, während der Fokus auf der Erhebung von *evaluative judgments* lag, ungeachtet dessen, dass *predictive judgments* einen integralen Bestandteil des Informationssuchprozesses darstellen (Lewandowski, 2008).

Neben Rieh (2002) unterscheiden auch Crystal & Greenberg (2006) und Tang & Solomon (2001) in ihren Studien explizit zwischen *predictive judgments* und *evaluative judgments*. Tang & Solomon (2001) legten beispielsweise für ihre Studie zur Wichtigkeit und Häufigkeit verwendeter Relevanzkriterien 90 Studierenden ein Surrogat inkl. Abstract und anschließend das Volltextdokument vor. Die Teilnehmenden sollten die von ihnen verwendeten Kriterien für die Auswahl der Volltexte nach deren Wichtigkeit beurteilen. Die Ergebnisse deuten darauf hin, dass die Relevanzbewertungen je nach Untersuchungsgegenstand (Abstract oder Volltext) auf unterschiedlich gewichteten Kriterien beruhen.

Tombros et al. (2005) und Savolainen & Kari (2006) fokussierten hingegen die Kriterien, die Nutzerinnen und Nutzer im Rahmen von *evaluative judgments* anwenden, im Gegensatz zu Balatsoukas & Ruthven (2012), die ausschließlich auf die Erhebung von *predictive judgments* abzielten. Für ihre Eye-Tracking-Studie gab es zwar weder Vorgaben hinsichtlich der Suchanfragen oder Informationsbedürfnisse noch zu dem verwendeten Suchsystem; ihre Analyse bezieht sich allerdings auf Suchergebnisse in Google. Die Autoren fanden einen Zusammenhang zwischen der Wahrnehmung von Surrogatkomponenten (Titel, Abstract und URL) und der Verwendung von 12 Relevanzkriterien (Balatsoukas & Ruthven, 2012, S. 1741):

- *Topicality*
- *Scope*
- *User Background*
- *Quality*
- *Tangibility*
- *Resource Type*

- *Affectiveness*
- *Recency*
- *Ranking*
- *Serendipity*
- *Format*
- *Document Characteristics*

Anhand dieser Kriterien wird erneut das Problem der fehlenden definitorischen und konzeptuellen Abgrenzung deutlich, zum Beispiel werden *Ranking*, *Serendipity*, *User Background* und *Format* als Kriterien betrachtet. Andere Studien konnten aufzeigen, dass diverse Aspekte die Relevanzbewertung in irgendeiner Weise beeinflussen; dennoch erscheint es wenig angemessen, die genannten Kriterien mit *Topicality* und *Scope* derselben Ebene zuzuweisen, vielmehr scheinen Kriterien wie *Serendipity* eine andere Ebene bzw. Bedeutungsdimension darzustellen.

Surrogate wurden auch in den Studien von Wang (1994), Howard (1994) und Maglaughlin & Sonnenwald (2002) als Bewertungsgrundlage gewählt, wenngleich der Begriff *predictive judgments* nicht verwendet wurde. Jede dieser Studien bezieht sich auf den akademischen Kontext.

Howard (1994) untersuchte Kriterien, die für die Operationalisierung von Pertinenz eine Rolle spielen und schlussfolgerte, „topicality appears to be more salient than informativeness“ (S. 184).

Maglaughlin & Sonnenwald (2002) fanden mithilfe von Interviews mit 12 Studierenden und einer Inhaltsanalyse 29 Kriterien und kategorisierten diese in sechs Gruppen (*Abstract, Author, Content, Full Text, Journal, Participant*), wobei inhaltsbezogene Kriterien der Gruppe *Content* die am häufigsten genannten Kriterien waren. Ähnlich zu der Identifizierung der Kriterien bei Balatsoukas & Ruthven (2012) zeigt sich auch hier eine unerwünschte Vermischung von Bedeutungsebenen unterschiedlicher Kriterien, wie beispielsweise Kriterien der Gruppe *Participant* im Vergleich mit den anderen Gruppen.

Die Arbeit von Wang (1994) wird im vorliegenden Abschnitt ausführlicher als andere Studien beleuchtet, da sie zwei wichtige Aspekte vereint: (1) Die Elemente eines Surrogats werden explizit erforscht und (2) die Auswahl von Dokumenten wird als ein Entscheidungsprozess basierend auf Kriterien und Merkmalen dargestellt, in welchem der Prozess der Relevanzbewertung mitinbegriffen ist. Diese Abgrenzung der Dokumentenauswahl von der Relevanzbewertung verweist auf den Unterschied zwischen Bewerten und Entscheiden; sie stellt eine wichtige Annahme und Voraussetzung für die experimentelle Erforschung von Relevanzkriterien dar und wird in Abschnitt 3.1.3 erneut aufgegriffen.

Für ihre Studie ließ Wang von 25 Teilnehmenden (Professoren und Professorinnen sowie Studierende der Agrarökonomie) Suchergebnisse aus DIALOG bewerten, nachdem die Teilnehmenden vorab ihre eigenen wissenschaftlichen Informationsbedürfnisse der Studienleitung im Rahmen von Interviews kommunizierten. Anhand der Informationsbedürfnisse wurden die Surrogate ermittelt und den Teilnehmenden zur Bewertung und Auswahl vorgelegt. Während dieses Prozesses wurden mithilfe der Methode des lauten Denkens[21] Aufnahmen erstellt, die für die Analyse herangezogen wurden. Das Ziel der Studie bestand unter anderem darin herauszufinden, welche Kriterien zur Bewertung bzw. Dokumentenauswahl herangezogen werden und welche Elemente eines Surrogats dabei eine Rolle spielen. Die Elemente eines Surrogats bezeichnet Wang (1994) als Dokumentinformationselemente (*document information elements – DIEs*), die sie folgender Einteilung unterzieht:

> *Descriptive DIEs* give clues to topicality, orientation, subject area, novelty, and recency, which can be interpreted straightforwardly. These DIEs are title, abstract, descriptors, geographical location, and publication date. *Inferential DIEs*, however, were used to judge orientation, quality, authority, and relation/origin, and, occasionally, topicality and subject area. Their interpretations depend on the users' personal knowledge and situation. Main inferential DIEs are author, author's affiliation, journal, and document type. Some DIEs, such as author and journal, may serve as both descriptive and inferential information elements. More-experienced users tended to use both types of DIEs to judge the documents. Less-experienced users tended to use only descriptive DIEs and occasionally inferential DIEs. (Wang, 1994, S. 190–191; Kursivdruck im Original nicht enthalten)

Von den Kriterien, welche die ableitbaren Elemente (*inferential DIEs*) umfassen, stellte sich Autorität als ein Kriterium heraus, das von erfahreneren Teilnehmenden angewendet wurde; als weiteres Ergebnis wurde festgestellt, dass Titel und Abstract die zwei wichtigsten Elemente darstellen und dass anhand dieser die thematische Relevanz (*topicality*) hauptsächlich bewertet wurde. Allerdings ist aufgrund der geringen Anzahl an Untersuchungspersonen ($n = 25$) und der Homogenität dieser Gruppe hinsichtlich des fachlichen Hintergrunds (Professorinnen/Professoren und Studierende aus dem Department Agrarökonomie) nicht davon auszugehen, dass es sich diesbezüglich um eine allgemeingültige Aussage handeln kann. Vor dem Hintergrund entscheidungstheoretischer Erkenntnisse entwickelte sie ein Modell, welches den kognitiven Prozess der Dokumentauswahl ab der Sichtung der Suchergebnisse bis zur Entscheidung darüber, ob das Surrogat ausgewählt wird, darstellt (Abbildung 2.3).

[21] Diese Methode wird in Abschnitt 2.2.1 näher erläutert.

Für die Entwicklung ihres Modells bezieht sich Wang (1994) unter anderem auf das von dem Psychologen Egon Brunswik (1952) entwickelte Linsenmodell (*lens model*) zur Entscheidungsfindung (auch *cue theory*). Das Linsenmodell besagt, dass Menschen zur Beurteilung eines Objekts, einer Variablen oder eines Kriteriums verschiedene Hinweise (*cues*) heranziehen und diese Informationen aggregiert betrachten – ähnlich wie eine optische Linse, die Licht bündelt, werden auch die Hinweisreize gebündelt. Wang (1994) nennt als Beispiele: „For example, disease (as a distal object or variable or criterion) can be judged by a series of symptoms (cues). Weather can be predicted by temperature, wind speed, and barometric pressure" (S. 28). Im Prozess der Dokumentenauswahl sind demnach die Elemente des Surrogats die vom Menschen wahrgenommenen Hinweisreize, welche die Basis für die Auswahlentscheidung bilden. Diese Elemente werden unter Anwendung von verschiedenen Kriterien kognitiv verarbeitet und interpretiert. Dabei wird das individuelle Wissen herangezogen, z. B. über das Fachgebiet, Personen (Autor-, Herausgeberschaft), Organisationen (Körperschaft, Verlag), das Publikationsorgan (Zeitschrift).

Wang (1994) legt in ihrem Modell elf Kriterien zugrunde, die in vorangegangen Studien, unter anderem von Linda Schamber und Carol Barry, ermittelt wurden (vgl. Abschnitt 2.1.1). Die Kriterien werden genutzt, um dem Dokument einen Wert zuzuweisen. Diese Zuweisung erfolgt anhand von Werten, die aus der volkswirtschaftlichen Theorie über das Konsumverhalten (*consumer choice theory*) übernommen sind. Die Theorie beruht auf der Annahme, dass das Kaufverhalten von fünf Werten (*consumption values*) beeinflusst wird: *functional value, conditional value, social value, emotional value* und *epistemic value* (Sheth et al., 1991, S. 160). Diese Werte überträgt Wang auf den Prozess der Dokumentenauswahl. Sie argumentiert, dass informationssuchende Personen – in Analogie zu Menschen, die ein Konsumgut wählen – ein Dokument auswählen, wenn sie es als wertvoll im Sinne von nützlich empfinden. Die Nützlichkeit (*utility*) eines Dokuments definiert Wang im Rahmen ihrer Studie als „the potential that the document has to satisfy an information need as perceived by the user in the situation" (Wang, 1994, S. 43). Nachfolgend sind die fünf Arten des Dokumentenwertes (*document value/worth*) aufgelistet (Wang & Soergel, 1998, S. 121–122):

- *Epistemic value* – the perceived utility of a document to satisfy a desire for knowledge or information that is unknown […]
- *Functional value* – the perceived utility of a document to make a contribution to the specific task at hand […]
- *Conditional value* – the perceived utility of a document is yet to be decided circumstantially […]

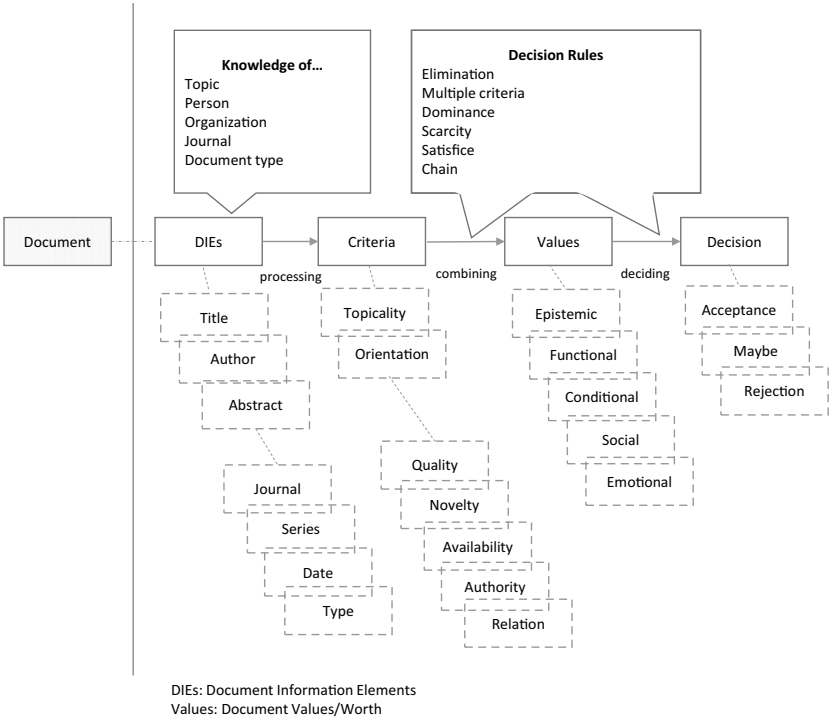

Abbildung 2.3 Kognitives Modell zur Dokumentenauswahl (nach Wang & Soergel, 1998)

- *Social value* – the perceived utility of a document in association with specific social groups or with individuals such as academic advisor, famous figures in the field, etc. […]
- *Emotional value* – the perceived utility of a document stemming from its capacity to arouse feelings or affective states […]

Im Anschluss an die Bewertung der Dokumente erfolgt die Entscheidung in drei Ausprägungen: (1) das Dokument wird akzeptiert, also ausgewählt, (2) das Dokument wird vielleicht zu einem späteren Zeitpunkt akzeptiert, es besteht Unsicherheit, (3) das Dokument wird abgelehnt, also nicht ausgewählt. Diese Entscheidungsprozesse werden von bestimmten Regeln geleitet (Montgomery, 1983;

Svenson, 1979), mit dem Ziel, kognitiven Aufwand zu verringern. Für den Prozess der Dokumentenauswahl identifizierte Wang auf Basis bestehender Literatur sechs Entscheidungsregeln (Wang & Soergel, 1998, S. 127–128):

1. *Elimination rule* – to reject a document, the user looks for an aspect of the document that enables him/her to quickly reject a document […]
2. *Multiple-criteria rule* – as a contrast to elimination rule, the user applies several criteria to accept or reject a document […]
3. *Dominance rule* – of similar documents, the user selects the one document which excels in at least one aspect and is not worse in the other aspects […]
4. *Scarcity rule* – when the user wants more documents, but only a few are retrieved, he/she tends to apply less stringent criteria so that even marginal documents are accepted […]
5. *Satisfice rule* – when the user feels that enough documents on a topic or facet have been selected, he/she may stop accepting relevant documents or terminate the selection process on that topic or facet […]
6. *Chain rule* – when the user identifies documents that are on a special chain, he/she tends to make a collective decision on the set […]

Für die Beschreibung der Entscheidungsregeln greift Wang auf ein volkswirtschaftliches Modell (*consumer choice theory*) zurück, das davon ausgeht, dass Menschen ihre Entscheidungen logisch und rational treffen. Wang selbst weist darauf hin, dass diese Annahmen nicht der Realität entsprechen, wie insbesondere durch die Arbeiten von Tversky und Kahnemann (z. B. Tversky & Kahneman, 1974) gezeigt werden konnte.

Gleichwohl ist Wangs Arbeit für die hier beschriebene Forschung von besonderer Bedeutung, da die gezielte Betrachtung der Elemente eines Surrogats als Grundlage der vorgenommenen Bewertung und ihre Wahrnehmung für die Anwendung von Kriterien im Kontext akademischer Suchsysteme wesentlich ist für die Identifikation der Einflüsse auf die Relevanzbewertungen und ihre definitorische Abgrenzung, worauf Kapitel 3 im Rahmen der Voraussetzungen zur experimentellen Erforschung von Relevanzkriterien abzielt. Ferner zeigt sich anhand von Wangs Untersuchung ein besonderer Punkt, der für die Einordnung früherer Studien zu Relevanzkriterien und für die Erforschung von Relevanzkriterien anhand von *predictive judgments* zentral ist:

In den frühen Studien enthielten die zu bewertenden Surrogate die klassischen Metadaten, wie sie beispielsweise in der Studie von Wang (1994) den Teilnehmenden vorgelegt wurden. Abbildung 2.4 zeigt exemplarisch eines der Surrogate aus der Studie aus einer DIALOG-Online-Datenbank, das vermutlich aus dem Jahr 1992 stammt, da die Datenerhebung im Sommer 1992 erfolgte (Wang &

Soergel, 1998, S. 119). Inzwischen hat sich die Gestaltung der Suchergebnisse und Suchinterfaces von Websuchmaschinen und anderen IR-Systemen nicht nur in Hinblick auf das Design und die Nutzerfreundlichkeit geändert. So sind heutzutage Popularitätsdaten wie die Anzahl der Zitationen und die Anzahl der Downloads eines Werkes ein integraler Bestandteil der Suchergebnisdarstellung in diversen akademischen Suchsystemen, wie beispielsweise in der fachübergreifenden Suchmaschine Google Scholar[22] (Abbildung 2.5), die bereits seit ihrer Einführung im November 2004 die Angabe „Cited by" anbietet (Jacsó, 2005), oder in der fachspezifischen Digitalen Bibliothek der *Association for Computing Machinery* (ACM Digital Library)[23] (siehe Abbildung 2.6, Stand 2017 bzw. Abbildung 2.7, Stand 2020).

Exemplarisch für die klassische Darstellung von Surrogaten in wissenschaftlichen Bibliothekskatalogen zeigt Abbildung 2.8 einen Screenshot des Online-Katalogs[24] der Universitätsbibliothek Hildesheim; zusätzliche Daten wie Popularitätsdaten sind hier nicht integriert. Die Überprüfung einiger Kataloge wissenschaftlicher Bibliotheken in Deutschland durch die Autorin zeigt, dass diese zum Zeitpunkt der Erfassung des Forschungsstands keine Popularitätsdaten in die Suchergebnispräsentation ihrer Suchsysteme integriert haben.

Auch in jüngeren Studien zu Relevanzkriterien, welche *predictive judgments* im akademischen Kontext erhoben, enthielten die Surrogate keine Popularitätsdaten. Eine Ausnahme bildet das von Bruza & Chang (2014) durchgeführte Experiment zur menschlichen Wahrnehmung der Relevanz von Dokumenten anhand von Surrogaten aus Google Scholar, inklusive der Anzahl an Zitationen als Popularitätsdatum. Die Autoren bezeichnen Relevanzkriterien als Dimensionen (*dimensions*) von Relevanz und führen beispielhaft die von Barry & Schamber (1998) identifizierten Kriterien an. Für ihr Experiment legten sie die folgenden sechs Dimensionen fest, von denen sie jeweils zwei anhand einer 4-Punkte-Skala durch Jurorinnen und Juroren pro Surrogat bewerten ließen: *Topicality, Credibility, Understandability, Believability, Interest, Sentimentality* (Bruza & Chang, 2014, S. 4,5). Die Ergebnisse implizieren, dass die Wahrnehmung der Relevanz von Snippets die kognitive Verarbeitung einer Vielzahl von Faktoren beinhaltet, zu denen die untersuchten Dimensionen (Kriterien) ebenfalls zählen. Welche Rolle einzelne Surrogatelemente, insbesondere die Anzahl der Zitationen, bei dieser kognitiven Verarbeitung spielen, wurde in dem Experiment weder gezielt untersucht noch ihr möglicher Einfluss auf die Relevanzwahrnehmung anerkannt.

[22] https://scholar.google.de/

[23] https://dl.acm.org/

[24] https://opac.lbs-hildesheim.gbv.de/

Citation # 16/5/1
TITLE: The On-Farm Costs of Reducing Groundwater Pollution
AUTHOR(S): Johnson, Scott L.; Adams, Richard M.; Perry, Gregory M.
AUTHOR(S) AFFILIATION: S Coast Air Quality Management District; OR State U ; OR State U
JOURNAL NAME: American Journal of Agricultural Economics,
JOURNAL VOLUME & ISSUE: 73 4, PAGES: 1063-73
PUBLICATION DATE: November 1991
DOCUMENT TYPE: Journal Article
ABSTRACT INDICATOR: Abstract
ABSTRACT: Agricultural chemicals are a source of groundwater pollution in some areas. Regulatory options to reduce such nonpoint pollution imply costs to producers. By integrating plant simulation, hydrologic, and economic models of farm-level processes, this study evaluates on-farm costs of strategies to reduce nitrate groundwater pollution. The empirical focus on intensively managed, irrigated farms in the Columbia Basin of Oregon. Results suggest that changes in timing and application rates of nitrogen and water reduce nitrate pollution with little loss in profits. Once such practices are adopted, further reductions in nitrates can be achieved only at increasing costs to producers.
GEOGRAPHIC LOCATION DESCRIPTOR(S): U.S.
DESCRIPTOR(S) (1991 forward only): Renewable Resources and Conservation; Environmental Management: Water; Air (Q250)
DESCRIPTOR(S): Conservation and Pollution (7220); Natural Resources--General (7210)

Abbildung 2.4 Surrogat aus DIALOG zu Beginn der 1990er Jahre (Wang, 1994, S. 85)

Information foraging in **information** access environments
P Pirolli, S Card - Proceedings of the SIGCHI conference on Human ..., 1995 - dl.acm.org
... ABSTRACT **Information foraging theory** is an approach to the analysis of human activities involving **information** access technologies. The ... hierarchy. KEYWORDS: **Information foraging theory**, **information** access. INTRODUCTION ...
☆ 99 Cited by 483 Related articles All 13 versions

Abbildung 2.5 Surrogat aus Google Scholar aus dem Jahr 2017

2.1.5 Zusammenfassung

Bei der Suche nach relevanten Informationsobjekten in Suchsystemen wenden informationssuchende Personen verschiedene Kriterien an, anhand derer sie die Relevanz von Suchergebnissen in Form von Surrogaten oder Volltexten bewerten. Jahrzehntelange informationswissenschaftliche Relevanzforschung hat eine Fülle an verschiedenen Einflussfaktoren und Relevanzkriterien ermittelt. Dabei fällt auf,

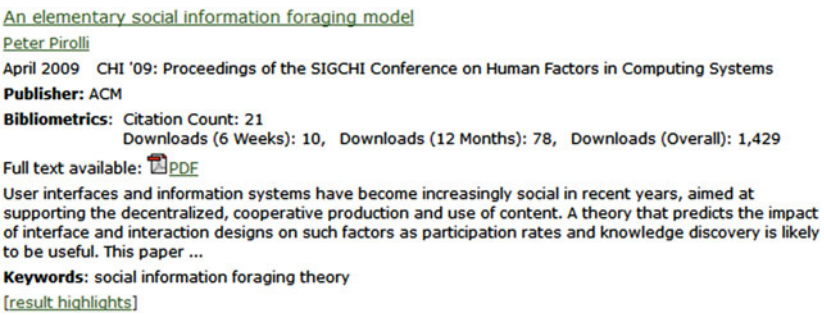
An elementary social information foraging model
Peter Pirolli
April 2009 CHI '09: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems
Publisher: ACM
Bibliometrics: Citation Count: 21
Downloads (6 Weeks): 10, Downloads (12 Months): 78, Downloads (Overall): 1,429
Full text available: PDF
User interfaces and information systems have become increasingly social in recent years, aimed at supporting the decentralized, cooperative production and use of content. A theory that predicts the impact of interface and interaction designs on such factors as participation rates and knowledge discovery is likely to be useful. This paper ...
Keywords: social information foraging theory
[result highlights]

Abbildung 2.6 Surrogat aus der ACM Digital Library aus dem Jahr 2017

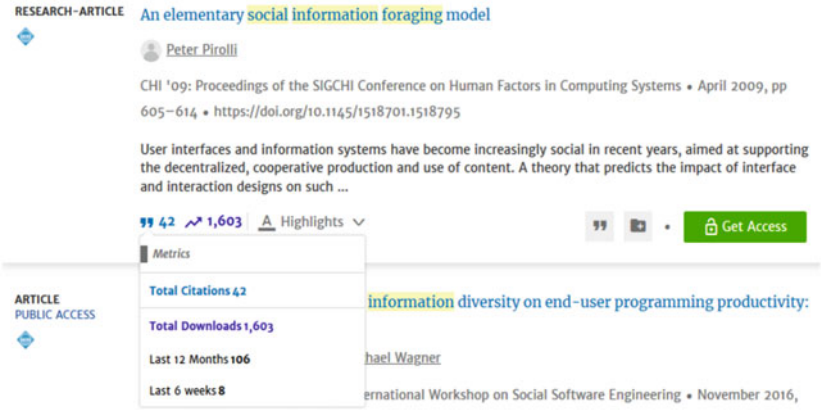
RESEARCH-ARTICLE
An elementary social information foraging model
Peter Pirolli
CHI '09: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems • April 2009, pp 605–614 • https://doi.org/10.1145/1518701.1518795
User interfaces and information systems have become increasingly social in recent years, aimed at supporting the decentralized, cooperative production and use of content. A theory that predicts the impact of interface and interaction designs on such ...
42 1,603 Highlights
Get Access
Metrics
Total Citations 42
Total Downloads 1,603
Last 12 Months 106
Last 6 weeks 8
ARTICLE
PUBLIC ACCESS
information diversity on end-user programming productivity:
hael Wagner
ernational Workshop on Social Software Engineering • November 2016,

Abbildung 2.7 Surrogat aus der ACM Digital Library aus dem Jahr 2020

dass die Bezeichnungen der Kriterien und Faktoren zum Teil sehr verwirrend sind (Bales & Wang, 2006; Maglaughlin & Sonnenwald, 2002; Wang, 2010; Xu & Chen, 2006) und durch die synonyme Verwendung von Kriterienbegriffen, die jedoch auf demselben Konzept beruhen, wie beispielsweise *currency, recency, novelty, timeliness* (Maglaughlin & Sonnenwald, 2002), eine definitorische und konzeptuelle Abgrenzung der beiden Begriffe Kriterium und Faktor erschwert wird.

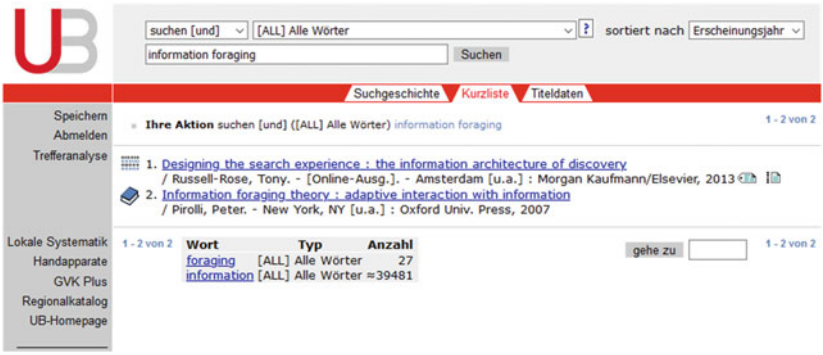

Abbildung 2.8 Suchergebnisdarstellung im Online-Katalog der Universitätsbibliothek Hildesheim aus dem Jahr 2020

Studien zu Relevanzkriterien stellten übereinstimmend die thematische Relevanz als Basiskriterium fest, die anhand der *Aboutness* von Dokumenten abgeleitet werden kann. Darauf aufbauend kommen weitere Kriterien hinzu, wie beispielsweise Validität, Aktualität und Glaubwürdigkeit. Letztere nimmt insbesondere im Webkontext im Zusammenhang mit Autorität bzw. kognitiver Autorität einen besonderen Stellenwert ein, da Informationen und Inhalte im Web vor ihrer Veröffentlichung nicht immer eine Qualitätskontrolle durchlaufen, wie dies im akademischen Kontext beispielsweise im Rahmen des Peer Review bei Manuskripten vor der Publikation oder bei Forschungsanträgen der Fall ist. Qualität taucht ebenfalls in der Literatur auf – manchmal als beeinflussender Faktor und in anderen Studien als Kriterium bei der Relevanzbewertung – obwohl Qualität als Oberbegriff oder Kriterienkategorie zu verstehen ist. So können Glaubwürdigkeit und Vertrauenswürdigkeit als Indikatoren für die Qualität eines Dokuments dienen, die beispielsweise mithilfe von Popularität als Rankingfaktor in Websuchmaschinen ermittelt werden soll.

Das Konzept der Popularität beruht auf dem Prinzip der Weisheit der vielen und findet auch Anwendung in linkbasierten Ranking-Verfahren (z. B. PageRank). Nutzersignale wie die Anzahl von Klicks oder auch die Verweildauer stellen Popularitätsdaten dar, die ebenfalls in das Ranking miteinfließen, jedoch nicht als sichtbarer Bestandteil in die Suchergebnispräsentation integriert sind und dennoch die Relevanzbewertung beeinflussen. Allerdings gibt es Popularitätsdaten, die als sichtbarer Bestandteil in die Ergebnispräsentation moderner akademischer

Suchsysteme integriert werden, wie die Anzahl von Zitationen bei Ergebnissen in Google Scholar oder zusätzlich die Downloadhäufigkeit eines Artikels in der ACM Digital Library. Demzufolge sind Surrogate mit zusätzlichen Daten wie Popularitätsdaten neben den erschließungstypischen Metadaten (Titel, Autor, Quelle, Erscheinungsjahr) als Grundlage der Relevanzbewertung von besonderem Interesse in Hinblick auf den möglichen Einfluss von Popularitätsdaten auf die Bewertung.

Die Ergebnisse früherer Studien zu Relevanzkriterien mit erhobenen Relevanzbewertungen auf der Basis von Surrogaten, also mithilfe von *predictive judgments* im Gegensatz zu *evaluative judgments* auf der Basis des vollständigen Inhalts, deuten darauf hin, dass Titel und Abstracts die meisten Hinweise für die Relevanzbewertung (Relevanzmerkmale) liefern. Diese früheren empirischen Studien zu Relevanzkriterien beruhen auf Surrogaten als Bewertungsgrundlage, die zu dem damaligen Zeitpunkt – wie es der Realität entsprach – keine Popularitätsdaten enthielten.

Aus diesem Grund stellt sich die Frage, inwieweit sich die Erkenntnisse zu den Kriterien, die Nutzerinnen und Nutzer akademischer Suchsysteme bei der Bewertung von Surrogaten anwenden, auf heutige Suchsysteme übertragen lassen. Bisher gibt es keine veröffentlichten Studien zu Relevanzkriterien auf der Basis von Surrogaten in akademischen Suchsystemen, die neben der Anzeige erschließungstypischer Metadaten mit zusätzlichen Daten wie Popularitätsdaten als (potenzielle) Relevanzmerkmale angereichert sind. Ob die Erkenntnisse aus diesen damaligen Studien zu Relevanzkriterien bei der Bewertung von Surrogaten im akademischen Kontext immer noch Gültigkeit besitzen, ist demzufolge unklar, wenngleich weiterhin von der thematischen Relevanz als Basiskriterium auszugehen ist.

2.2 Methoden zur Erforschung von Relevanzkriterien

Die vorliegende Arbeit beschäftigt sich mit den Kriterien, die informationssuchende Personen während der Interaktion mit einem Suchsystem der Relevanzbewertung anwenden. Relevanzkriterien werden dabei als rein subjektiv erachtet (vgl. 2.1). Das bedeutet, dass Relevanzkriterien nicht unabhängig vom Menschen untersucht werden können. Für die Erforschung von Relevanzkriterien ist die Verwendung sozialwissenschaftlicher Methoden sinnvoll, da sozialwissenschaftliche Methoden generell verwendet werden, um Sachverhalte, die das Erleben und Verhalten von Menschen betreffen, zu ergründen (Döring & Bortz, 2016, S. 4).

Insbesondere die sozialwissenschaftliche Fachdisziplin der Psychologie erforscht menschliches Erleben und Verhalten mithilfe von empirischen Studien.

Zu unterscheiden sind bei empirischen Studien die Art des Untersuchungsdesigns und die Art der Datenerhebung. Bezüglich ihres Erkenntnisinteresses lassen sich empirische Studien als deskriptive, explorative und explanative Untersuchungsdesigns kategorisieren (Döring & Bortz, 2016, S. 192). Sozialwissenschaftliche Methoden der Datenerhebung lassen sich in fünf Gruppen unterteilen: (a) Befragung, (b) Beobachtung, (c) psychologischer Test, (d) physiologische Messungen und (e) Dokumentenanalyse. Bei allen Methoden kommen sowohl qualitative als auch quantitative Verfahren zum Einsatz, nur bei der physiologischen Messung gibt es ausschließlich quantitative Verfahren (Döring & Bortz, 2016, S. 312–577). Im Mittelpunkt stehen die Verfahren der Befragung und Beobachtung, die auch in der bibliotheks- und informationswissenschaftlichen Forschung am häufigsten Anwendung finden (Connaway & Radford, 2017, S. 17–19).

Die nachfolgenden Abschnitte betrachten Studien zu Relevanzkriterien aus methodischer Perspektive. Unterschieden werden diese nach der Art der Datenerhebung in Befragungen (Abschnitt 2.2.1) und Beobachtungen (Abschnitt 2.2.2) sowie nach der Art des Untersuchungsdesigns[25] in explorative (Abschnitt 2.2.3) und experimentelle[26] (Abschnitt 2.2.4) Studien. Die Erkenntnisse beruhen auf der Analyse von 47 informationswissenschaftlichen Studien[27] zu Relevanzkriterien. Ausgehend von publizierten Literaturschauen und den darin zitierten Quellen (z. B. Mizzaro, 1997; Saracevic, 2016b) wurden mittels *backward chaining* und *forward chaining* (Ellis, 1989) Studien, die zwischen 1988 und 2016 erschienen, für die Analyse ausgewählt. Da die Literaturschauen nur englischsprachige Quellen, insbesondere aus dem angloamerikanischen Raum, beinhalten, ist die Auswahl der Studien für die hier vorgestellte Analyse ebenfalls auf englischsprachige Publikationen beschränkt.

---

[25] Deskriptive Studien, die eine Population beschreiben, sind bei informationswissenschaftlichen Studien zu Relevanzkriterien nicht zu finden; daher wird auf einen entsprechenden Abschnitt verzichtet.

[26] Zwar wäre die Bezeichnung *explanative Untersuchungsdesigns* korrekt; sie wird für den Abschnitt jedoch nicht verwendet, weil in diesem der Fokus auf experimentelle Studien als solche, die unter den hypothesentestenden, vollstrukturierten Studien „[d]ie höchste Erklärungskraft im Hinblick auf den Nachweis von Kausalität [besitzen]“ (Döring & Bortz, 2016, S. 192), gelegt wird.

[27] Abschnitt 2.1 betrachtet Studien in Hinblick auf deren inhaltliche Erkenntnisse zu Relevanzkriterien, aber nicht alle dort zitierten Studien sind in der Analyse enthalten. Eine Übersicht der 47 analysierten Studien ist in Anhang 1 im elektronischen Zusatzmaterial enthalten.

Die Auswahl der Studien erfolgte anhand zweier grundsätzlicher Kriterien, die beide erfüllt sein mussten: (1) Die Studie beschreibt (unter anderem) Kriterien, die Nutzerinnen und Nutzer während der Bewertung oder Auswahl von Dokumenten anwenden; (2) es handelt sich um eine empirische Studie, die auf den von menschlichen Nutzerinnen und Nutzern erhobenen Daten beruht. Zusätzlich zur Art des Forschungsdesigns und der Methode der Datenerhebung bestand ein Interesse an weiteren methodischen Merkmalen, die ein besseres Bild über das Vorgehen bei der Erforschung von Relevanzkriterien erlauben. So wurden die 47 ausgewählten Studien systematisch erfasst hinsichtlich:

(a) der verwendeten Methode zur primären Datenerhebung;
(b) der Art des Untersuchungsdesigns;
(c) der Anzahl der Teilnehmenden;
(d) der Frage, ob die Relevanzkriterien gegeneinander gewichtet wurden;
(e) der Art der Skala für die Erfassung expliziter Relevanzbewertungen;
(f) des Kontexts der Suchaufgaben (z. B. akademisch, schulisch, Alltag[28]).

Die Anzahl der Studienteilnehmenden, also der Stichprobenumfang, gibt Aufschluss über die Aussagekraft der Ergebnisse, wobei zum einen eine größere Anzahl von Teilnehmenden nicht zwangsläufig zu besseren Ergebnissen führt und zum anderen die erforderliche Stichprobengröße von weiteren Parametern abhängt.[29] Informationen darüber, ob die ermittelten Relevanzkriterien gegeneinander gewichtet wurden, sind erforderlich in Hinblick auf die Wichtigkeit oder Nachrangigkeit verwendeter Kriterien. Zusätzlich wurde der Kontext der Suchaufgaben erfasst, da diese Information für die inhaltliche Einordnung der Kriterien und Eingrenzung der Studien erforderlich war (vgl. Abschnitt 2.1).

Die Analyse ergab, dass in den häufigsten durchgeführten Studien ($n = 21$) die Teilnehmenden Suchaufgaben im akademischen Kontext bearbeiteten. Abbildung 2.9 veranschaulicht die Häufigkeiten der Studien im akademischen Kontext (AKAD), im Alltagskontext (ELIS), im Arbeitskontext bzw. beruflichen Umfeld (ARB), im schulischen Kontext (SCHUL) sowie übergreifend (AKAD/ELIS bzw. ARB/ELIS); in drei Studien wurde der Kontext weder genannt, noch konnte er aus der Beschreibung abgeleitet werden (N. a.).

[28] Der Kontext Alltag bezieht sich auf die Art von Informationssuche, die in der Informationswissenschaft als *Everyday Life Information Seeking* (ELIS) bezeichnet wird.

[29] Erläuterungen zur Berechnung der erforderlichen Anzahl an Untersuchungspersonen für eine ausreichende Teststärke erfolgen im Zusammenhang mit der Beschreibung des im Rahmen dieser Arbeit durchgeführten Experiments in Abschnitt 4.2.4.

Abbildung 2.10 gibt einen Überblick über die Stichprobengrößen der einzelnen Studien, die von 3 bis 350 reichen, wobei der Durchschnitt 48 Teilnehmende und der Median 24 betragen.

Anhand von Abbildung 2.11 lässt sich ein stetiges Interesse an der Erforschung von Relevanzkriterien in den letzten Jahrzehnten erkennen: Zwar wurden die bekanntesten Studien zu Relevanzkriterien (z. B. von Barry und Schamber, siehe Abschnitt 2.1.1) in den 1990er veröffentlicht, doch auch nach der Jahrtausendwende wurden Relevanzkriterien weiterhin erforscht, insbesondere im Kontext der Websuche (vgl. Abschnitt 2.1.2).

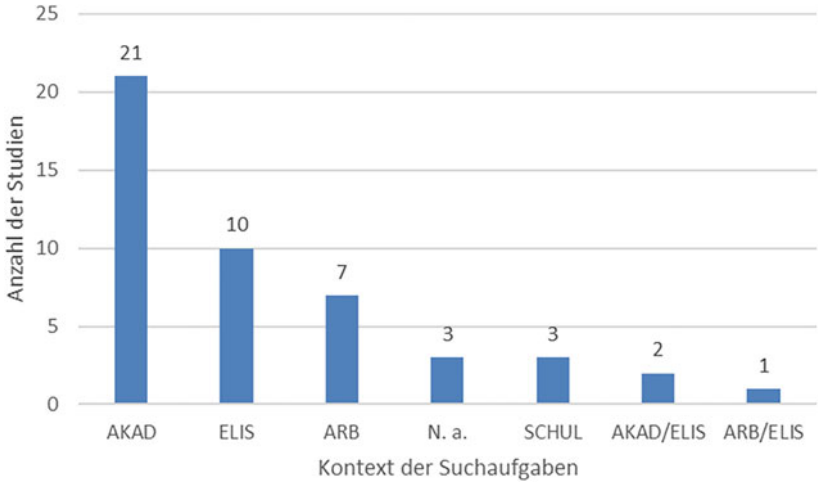

Abbildung 2.9 Anzahl der Studien nach Aufgabenkontext (N = 47)

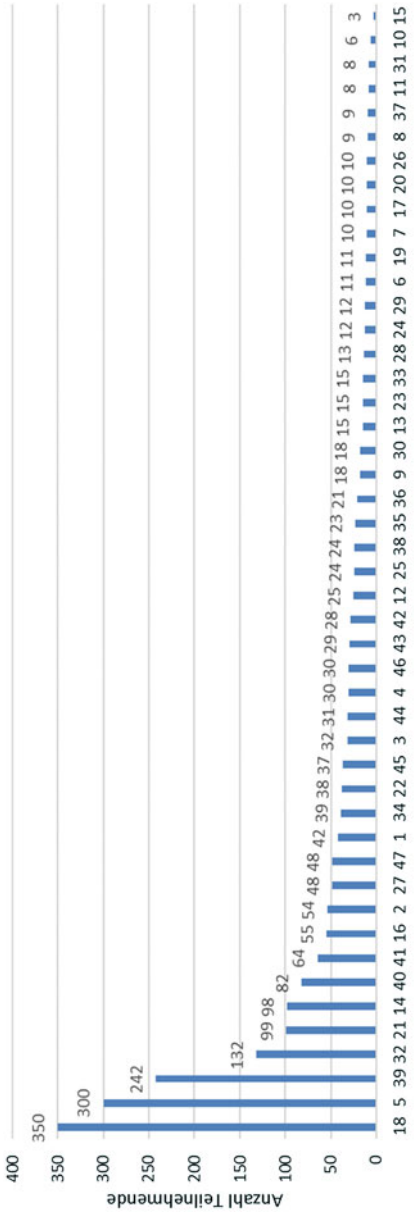

Abbildung 2.10 Anzahl der Teilnehmenden pro Studie (N = 47)

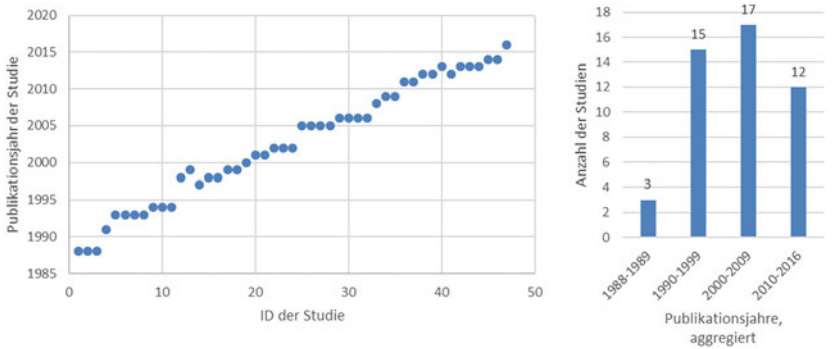

Abbildung 2.11 Publikationsjahre der 47 Studien, einzeln (links) und aggregiert (rechts)

2.2.1 Befragungen

Zur Befragung zählt die Erhebung mittels Fragebogen, welche sowohl schriftlich als auch mündlich erfolgen kann und auch als Umfrage (*survey*) oder Umfrageforschung (*survey research*) bezeichnet wird (Döring & Bortz, 2016, S. 356 ff.). Interviews stellen ebenso eine Befragungsart dar. Beide Befragungsmethoden werden nach dem Grad ihrer Strukturierung unterteilt in un- und halbstrukturierte (qualitative) sowie vollstrukturierte (quantitative) Verfahren. Als eine besondere Form des unstrukturierten Interviews wird die Methode des lauten Denkens (*think aloud*) gezählt, bei der die Testpersonen ihre Gedanken während einer Handlung verbalisieren. Dieser Vorgang wird aufgezeichnet, sodass ein Think-aloud-Protokoll entsteht, welches analog zu Interviewtranskripten von den Forschenden analysiert wird (Döring & Bortz, 2016, S. 371). Begründet wird die Zuordnung der Methode des lauten Denkens zu den Befragungsmethoden wie folgt:

> Die Methode des lauten Denkens ist insofern den mündlichen Befragungsmethoden zuzuordnen, als die Untersuchungspersonen ihre Gedanken in Worte fassen und mündliche verbale Daten generiert werden, die ohne den Forschungsprozess nicht existieren würden. Im Unterschied zu anderen qualitativen Interviewvarianten spielt jedoch beim lauten Denken die Interaktion zwischen Auskunftspersonen und Interviewenden keine Rolle, vielmehr ergeben sich die Äußerungen als Kommentare zu einem selbst absolvierten Handlungsablauf. (Döring & Bortz, 2016, S. 370–371).

Eine Variante der halbstrukturierten Befragung sind Diskussionen in Fokusgruppen. Im Gegensatz zum Einzelinterview können in Fokusgruppen unterschiedliche Meinungen und Aussagen der einzelnen Mitglieder hervortreten und innerhalb der Gruppe – zum Teil auch kontrovers – diskutiert werden, wobei das dabei im Fokus stehende Thema durch die Forschungsleitung vorgegeben wird (Döring & Bortz, 2016, S. 359). Die halbstrukturierte Tagebuchmethode stellt eine Form der schriftlichen Befragung dar und besteht in der Erhebung von Daten aus Tagebucheinträgen zu bestimmten Themen durch die befragten Personen (Döring & Bortz, 2016, S. 405). Dagegen handelt es sich bei der vollstrukturierten Tagebuchmethode um das Ausfüllen vollstandardisierter schriftlicher Fragebögen durch die Untersuchungspersonen (Döring & Bortz, 2016, S. 418).

Die Analyse der 47 ausgewählten Studien zu Relevanzkriterien ergab, dass die Befragung die meistverwendete Methode zur Datenerhebung darstellt, die in *jeder* der Studien vorgenommen wurde. In 30 Studien wurden Daten *ausschließlich* mittels Verfahren der Befragung erhoben (Tabelle 2.4). Fokusgruppeninterviews wurden lediglich in zwei der untersuchten Studien durchgeführt (vgl. Tabelle 2.4). So führten beispielsweise Walraven, Brand-Gruwel, & Boshuizen (2009) Fokusgruppeninterviews mit 23 Schülerinnen und Schüler der Sekundarstufe aus zwei Schulen durch, unter anderem mit dem Ziel, über deren Wissen und konzeptuelle Vorstellungen zu den Kriterien bei der Evaluierung von Suchergebnissen, Quellen und Informationen auf einer Webseite zu erfahren. Sie fanden beispielsweise heraus, dass die Kriterien zur Beurteilung von Suchergebnissen sich auf diese Elemente bezogen: (1) Titel/Snippet, (2) Art (Webseite/PDF), (3) URL/Domäne, (4) Position in der Ergebnisliste (Ranking), (5) Bekanntheit, (6) Sprache.

Ein großer Teil der Studien ($n = 20$) verwendet die Methode des lauten Denkens (Tabelle 2.5), die auch häufig in IIR-Studien verwendet wird; der Nachteil dieser Methode besteht darin, dass die Teilnehmenden oft Schwierigkeiten bei der Artikulation ihrer Gedanken während der Bearbeitung einer Suchaufgabe haben, sodass die Forschungsleitung nachfragen muss, was die Künstlichkeit der Situation vermutlich verstärkt (Kelly, 2009, S. 84 ff.). Vor diesem Hintergrund erscheint es fragwürdig, Kinder mithilfe dieser Methode zu befragen, um die von ihnen genutzten Kriterien bei der Suche nach Quellen für ein Schulprojekt herauszufinden, wie es in der Studie von Hirsh (1999) erfolgte. Allerdings zeigten die wenigsten der zehn Kinder im Alter von 10 bis 11 Jahren Probleme bei der Begründung ihrer ausgewählten Informationsobjekte; Jungen hatten weniger Schwierigkeiten bei der Verbalisierung der ihren Entscheidungen zugrunde liegenden Kriterien oder Faktoren als Mädchen (Hirsh, 1999, S. 1277), jedoch können insbesondere vor diesem Hintergrund und in Hinblick auf die sehr geringe Stichprobengröße keine allgemeingültigen Schlussfolgerungen getroffen werden.

Die Tagebuchmethode fand lediglich in drei der 47 Studien Anwendung (Tabelle 2.6). Watson (2014) bat 37 Schülerinnen und Schüler im Alter von 14 bis 17 Jahren, Tagebucheintragungen in Form eines Journals für die Bearbeitung ihrer Schulaufgaben vorzunehmen, führte Interviews mit ihnen durch und zeichnete Suchprozesse auf, die zusätzlich mithilfe der retrospektiven Methode des lauten Denkens durch die teilnehmenden Kinder kommentiert wurden. Die Auswertung folgte dem Grounded-Theory-Ansatz, der in der qualitativen Forschung häufig Anwendung findet. Das Forschungsziel bestand darin, die Kriterien zur Bewertung der Relevanz und Zuverlässigkeit von Informationsobjekten aufzudecken (Watson, 2014).

Tabelle 2.4 Studien mit Verfahren der Befragung (ohne Beobachtung)

ID	Quelle	Anzahl Testpersonen	Kriterien gewichtet?	Kontext
1	Halpern & Nilan (1988)	42	Nein	ELIS
2	Nilan, Peek, & Snyder (1988)	54	Nein	ELIS
3	Regazzi (1988)	32	Ja	AKAD
4	Schamber (1991)	30	Nein	ARB
5	Cool, Belkin, Kantor, & Frieder (1993), Studie 1	300	N. a.	AKAD
6	Cool, Belkin, Kantor, & Frieder (1993), Studie 2	11	N. a.	AKAD
7	Park (1993)	10	Nein	AKAD
8	Thomas (1993)[a]	9	Nein	N. a.
9	Barry (1994)	18	Nein	AKAD
10	Bruce (1994)	6	Ja	AKAD
11	Howard (1994)	8	N. a.	AKAD
12	Wang (1994); Wang & Soergel (1998)	25	Nein	AKAD
13	Wang & White (1995, 1999)	15	Nein	AKAD

(Fortsetzung)

Tabelle 2.4 (Fortsetzung)

ID	Quelle	Anzahl Testpersonen	Kriterien gewichtet?	Kontext
14	Fidel & Crandall (1997)	98	Nein	ARB
18	Schamber & Bateman (1999) & Bateman (1998, 1999)	350	Ja	N. a.
21	Tang (1999); Tang & Solomon (2001)	99	Ja	AKAD
22	Choi & Rasmussen (2002)	38	Ja	AKAD
24	Maglaughlin & Sonnenwald (2002)	12	Nein	AKAD
26	Hung, Zoeller, & Lyon (2005)	10	Nein	ARB
32	Xu & Chen (2006)	132	Nein	ELIS
34	Art Taylor, Zhang, & Amadio (2009)	39	Nein	AKAD
35	Walraven, Brand-Gruwel, & Boshuizen (2009)[a]	23	Nein	SCHUL
36	Beresi, et. al. (2010); Beresi (2011)	21	Nein	AKAD
37	Chu (2011)	9	Ja	AKAD
39	De Sabbata & Reichenbacher (2012)	242	Ja	ELIS
40	Taylor (2012a); Taylor (2013) Studie 1	82	Nein	AKAD
41	Taylor (2012b) Studie 2	64	Nein	AKAD
43	Sedghi, Sanderson, & Clough (2013)	29	Nein	ARB
45	Watson (2014)	37	Nein	SCHUL
46	Zhang (2014)	30	Nein	ELIS

[a] Studien, in denen Fokusgruppeninterviews durchgeführt wurden (n = 2)
N = 30

Tabelle 2.5 Studien mit der Think-aloud-Methode

ID	Quelle	Anzahl Testpersonen	Kriterien gewichtet?	Kontext
7	Park (1993)	10	Nein	AKAD
12	Wang (1994); Wang & Soergel (1998)	25	Nein	AKAD
14	Fidel & Crandall (1997)	98	Nein	ARB
17	Hirsh (1999)	10	Nein	SCHUL
19	Vakkari & Hakala (2000)	11	Nein	AKAD
20	Fitzgerald & Galloway (2001)	10	Nein	AKAD
21	Tang (1999); Tang & Solomon (2001)	99	Ja	AKAD
23	Rieh (2002)	15	Nein	AKAD/ELIS
25	Tombros, Ruthven, & Jose (2003, 2005)	24	Nein	ELIS
27	Toms, O'Brien, Kopak, & Freund (2005)	48	Nein	AKAD/ELIS
28	Twait (2005)	13	Nein	AKAD
29	Crystal & Greenberg (2006)	12	Nein	ELIS
30	Savolainen & Kari (2006)	18	Nein	ELIS
33	Papaeconomou, Zijlema, & Ingwersen (2008)	15	Nein	ELIS
35	Walraven, Brand-Gruwel, & Boshuizen (2009)	23	Nein	SCHUL
36	Beresi, et. al. (2010); Beresi (2011)	21	Nein	AKAD
38	Balatsoukas & Ruthven (2012)	24	Nein	N. a.
43	Sedghi, Sanderson, & Clough (2013)	29	Nein	ARB

(Fortsetzung)

Tabelle 2.5 (Fortsetzung)

ID	Quelle	Anzahl Testpersonen	Kriterien gewichtet?	Kontext
44	Xie, Benoit, & Zhang (2010); Xie & Benoit (2013)	31	Nein	ARB/ELIS
45	Watson (2014)	37	Nein	SCHUL

N = 20

Tabelle 2.6 Studien mit der Tagebuchmethode

ID	Quelle	Anzahl Testpersonen	Kriterien gewichtet?	Kontext
18	Bateman (1998, 1999); Schamber & Bateman (1999)	350	Ja	N. a.
19	Vakkari & Hakala (2000)	11	Nein	AKAD
45	Watson (2014)	37	Nein	SCHUL

N = 3

Die Erfassung von expliziten Relevanzbewertungen stellt ebenfalls eine Form der Befragung dar und es erscheint naheliegend zu erwarten, dass zur Erforschung von Relevanzkriterien auch Relevanzbewertungen erhoben werden. Dies ist jedoch bei den untersuchten Studien nicht immer der Fall. In nur 20 der 47 Studien wurden von den Teilnehmenden explizit vorgenommene Relevanzbewertungen erfasst und ausgewertet. Zumeist erfolgte die Erfassung mittels Rating-Skalen, wobei drei Kategorien (*nicht relevant – teilweise relevant – relevant* bzw. *nicht relevant – relevant – weiß nicht/bin unsicher*) am häufigsten verwendet wurden; binäre Relevanzbewertungen mithilfe einer dichotomen Skala wurden nur in zwei Studien erfasst; in ebenso wenigen wurde die Methode der *Magnitude Estimation*[30] angewandt (Abbildung 2.12).

Obgleich in 27 Studien auf die Erhebung expliziter Relevanzbewertungen verzichtet wurde, bedeutet dies nicht, dass keine Form der Bewertung stattfand. Auch mithilfe der Methode des lauten Denkens, bei der es um das Erfassen kognitiver Prozesse während der Wahrnehmung von Informationsobjekten geht, findet eine

[30] Auf diese Methode zur Erhebung expliziter Relevanzbewertungen wird im Zusammenhang mit der im Rahmen dieser Arbeit durchgeführten experimentellen Studie in Abschnitt 4.1.2 näher eingegangen.

Form von mündlicher, subjektiver Bewertung statt, die sich in den Erläuterungen (dem lauten Denken) der Teilnehmenden niederschlägt. Es kann demnach geschlussfolgert werden, dass sich Relevanzkriterien nicht unabhängig von einer Bewertung erforschen lassen, auch wenn die Bewertung implizit stattfindet bzw. nicht explizit erfasst wird.

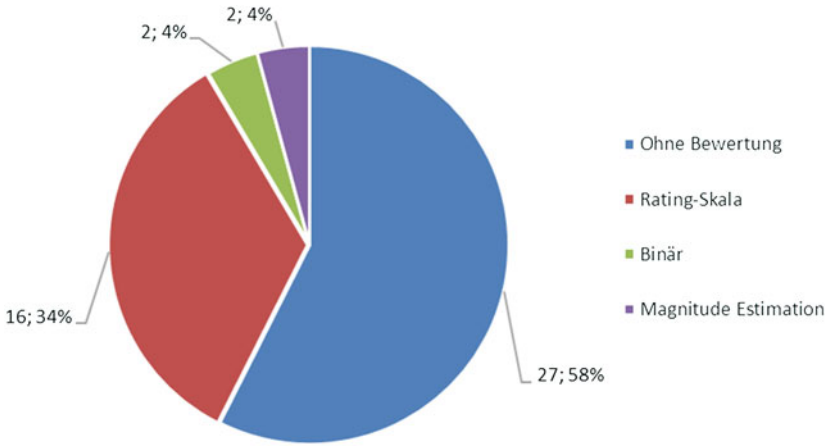

Abbildung 2.12 Anteil der Studien mit erhobenen Relevanzbewertungen (N = 47) und Art der Skalenerhebung

2.2.2 Beobachtungen

Während bei der Befragung verbale Äußerungen der Untersuchungspersonen in schriftlicher oder mündlicher Form erfasst werden, zielt die Beobachtung auf die „Dokumentation und Interpretation von Merkmalen, Ereignissen oder Verhaltensweisen mithilfe menschlicher Sinnesorgane und/oder technischer Sensoren zum Zeitpunkt ihres Auftretens“ (Döring & Bortz, 2016, S. 324). Qualitative Verfahren weisen dabei keine oder nur eine geringe Strukturierung auf (z. B. ethnografische Feldbeobachtung, autoethnografische Selbstbeobachtung); quantitative Verfahren der Beobachtung sind dagegen stark strukturiert und finden oft in einem Labor zur Erfassung von Verhaltensreaktionen statt (z. B. bei einer experimentellen Laborstudie). Das erhobene Datenmaterial liegt in Form von Beobachtungsprotokollen vor.

Werden zu Beobachtungszwecken technische Hilfsmittel oder Software eingesetzt, handelt es sich um technikvermittelte Beobachtungen, wie beispielsweise das Speichern von Klickdaten bei Online-Tests und Daten aus Logfiles (*transaction logs*). Die Erfassung mit technischen Hilfsmitteln bedeutet zudem, dass auch die *physiologische Messung* eine Form der quantitativen Beobachtung darstellt (Döring & Bortz, 2016, S. 523 ff.). Hierzu zählt beispielsweise das Messen von Blickbewegungen (*eye-tracking*), bei der mithilfe eines Eye-Trackers die Bewegungen der Pupillen der Untersuchungspersonen während einer Handlung (z. B. der Interaktion mit einer Suchmaschine) aufgezeichnet werden.

Von den 47 analysierten empirischen Studien zu Relevanzkriterien wurden in 17 zusätzlich zur Befragung Verfahren der Beobachtung angewendet. Von diesen bezogen sechs Studien die Auswertung von Klickdaten aus Logfiles mit ein, in zwei Studien wurden die Augenbewegungen der Testpersonen mithilfe einer Eye-Tracking-Software gemessen (Tabelle 2.7).

Logfile-Analysen werden in Interactive Information Retrieval-Studien häufig genutzt, um Interaktionen von Nutzerinnen und Nutzern mit einem IR-System aufzuzeichnen und dabei jedes Event (z. B. Eingaben, Klicks und Adressen der geklickten Objekte) nachverfolgen zu können (Kelly, 2009, S. 87 ff.). Diese Aufzeichnungen liefern zusätzliche Informationen wie Zeitstempel, anhand derer Erkenntnisse zum Informationssuchverhalten erzielt werden sollen. Die Grenze der Methode allgemein liegt darin, dass Klickdaten abgesehen von der übermittelten Suchanfrage keine näheren Informationen über den Kontext oder das konkrete Informationsbedürfnis bzw. das Informationsproblem der informationssuchenden Person liefern (Kelly, 2009, S. 91) und die Interpretation solcher Daten eine Herausforderung für Forschende bedeutet. Dies ist somit ein Grund, warum in jeder der 47 Studien zu Relevanzkriterien mindestens die Form der Befragung als Erhebungsmethode verwendet wurde.

Tabelle 2.7 Studien mit Verfahren der Beobachtung (zugleich Befragung)

ID	Quelle	Anzahl Testpersonen	Kriterien gewichtet?	Kontext
15	Markkula & Sormunen (1998)	3	Ja	ARB
16	Spink, Greisdorf, & Bateman (1998)[a]	55	Nein	AKAD
17	Hirsh (1999)	10	Nein	SCHUL
19	Vakkari & Hakala (2000)[a]	11	Nein	AKAD
20	Fitzgerald & Galloway (2001)	10	Nein	AKAD
23	Rieh (2002)[a]	15	Nein	AKAD/ELIS
25	Tombros, Ruthven, & Jose (2003, 2005)	24	Nein	ELIS
27	Toms, O'Brien, Kopak, & Freund (2005)[a]	48	Nein	AKAD/ELIS
28	Twait (2005)	13	Nein	AKAD
29	Crystal & Greenberg (2006)	12	Nein	ELIS
30	Savolainen & Kari (2006)	18	Nein	ELIS
31	Westman & Oittinen (2006)[a]	8	Ja	ARB
33	Papaeconomou, Zijlema, & Ingwersen (2008)[b]	15	Nein	ELIS
38	Balatsoukas & Ruthven (2012)[b]	24	Nein	N. a.
42	Kim, Kazai, & Zitouni (2013)	28	Nein	ELIS
44	Xie, Benoit, & Zhang (2010); Xie & Benoit (2013)[a]	31	Nein	ARB/ELIS

(Fortsetzung)

Tabelle 2.7 (Fortsetzung)

ID	Quelle	Anzahl Testpersonen	Kriterien gewichtet?	Kontext
47	Hamid, Thom, & Iskandar (2016)	48	Ja	ARB

[a] Studien, in denen Daten aus Logfiles erhoben wurden; n = 6.
[b] Studien, in denen Augenbewegungen gemessen wurden (Eye-Tracking); n = 2.
N = 17

Ein weiterer Punkt, den es im Zusammenhang mit Klickdaten zu berücksichtigen gilt, besteht bei der Erhebung von Relevanzbewertungen: Als Alternative zur Erhebung expliziter Bewertungen ist die Erfassung der Anzahl von Dokumenten, die angeklickt oder nicht angeklickt wurden, wenig sinnvoll, weil ein Klick ein binäres Verständnis von Relevanz[31] widerspiegelt. Vielmehr zeigt ein Klick eine Entscheidung für oder gegen die Auswahl eines Treffers an; dieser Entscheidung geht eine Beurteilung voraus, die als Ergebnis des Bewertungsprozesses[32] betrachtet werden kann.

Im Gegensatz zur Analyse von Klickdaten erlaubt das Messen von Blickbewegungen nachzuverfolgen, welche Elemente auf dem Bildschirm wie lange und in welcher Reihenfolge die Aufmerksamkeit der Testpersonen auf sich zogen. In Studien zur Erforschung des Informationssuchverhaltens oder in IR-Studien sind dabei die sakkadischen Blickbewegungen von Interesse, bei denen es sich um schnelle, ruckartige Bewegungen handelt, die typischerweise beim Lesen auftreten (Döring & Bortz, 2016, S. 522). Ein Nachteil bei diesem Messverfahren bestand bei den damaligen Studien darin, dass sich Eye-Tracking für die Testpersonen seltsam anfühlen konnte (Kelly, 2009, S. 199) und unter Umständen die ohnehin als wenig realistisch empfundene Situation in einem Testlabor auf die Teilnehmenden noch künstlicher wirkte. Moderne Eye-Tracker für den Desktop-Bereich sind verhältnismäßig wenig störend. Dennoch besteht der Nachteil dieser Art der Datenerhebung wie bei allen Studien, in deren Rahmen die Teilnehmenden ein Labor aufsuchen müssen, in der Herausforderung der Probandenakquise und dem Erreichen einer ausreichend großen Stichprobe.

In der Studie von Balatsoukas & Ruthven (2012) wurden 24 Teilnehmende bei ihrer Interaktion mit Google unter Nutzung eines Eye-Trackers beobachtet

[31] Tatsächlich handelt es sich bei dem Relevanzkonzept nicht um ein binäres Konzept, wie im Rahmen der Konzeptspezifikation in Abschnitt 3.1 dargestellt wird.

[32] Die Begründung, die Relevanzbewertung von Informationsobjekten als einen Prozess anzusehen, wird in Abschnitt 3.1.3 erläutert. Dabei wird auch auf die Unterscheidung von Bewerten (Urteilen) und Entscheiden ausführlicher eingegangen.

und zugleich der Methode des lauten Denkens unterzogen, um herauszufinden, welcher Zusammenhang zwischen der Anwendung von Relevanzkriterien und bestimmten Elementen der Suchergebnisse besteht. Mithilfe einer 3-Punkte-Rating-Skala wurden explizite Relevanzbewertungen von Surrogaten (*predictive judgments*) erhoben und eine Verbindung von Surrogat-Komponenten und Relevanzkriterien im Zusammenhang mit der Anzahl der Fixationen hergestellt. Die Ergebnisse zeigen unter anderem, dass im Durchschnitt die größte Anzahl an Fixationen für das Kriterium der thematischen Relevanz (*topicality*) auf Basis des Titels und des Abstracts aufgewendet wird (vgl. Abschnitt 2.1.4). Die Autoren merken an, dass die größte methodische Herausforderung bei der Datenanalyse darin bestand, eine Verbindung der Blickbewegungsdaten (Anzahl und Dauer der Fixationen) mit den qualitativen Daten (z. B. Arten von Relevanzkriterien, Relevanzstufen) herzustellen (Balatsoukas & Ruthven, 2012, S. 1743).

In den restlichen neun der 17 Studien, in denen mittels Beobachtungsverfahren Daten erhoben wurden, kamen folgende Beobachtungsverfahren zum Einsatz: Twait (2005) notierte ihre Beobachtungen von 13 Studierenden, während diese ihre Informationssuche mithilfe der Methode des lauten Denkens verbalisierten; Crystal & Greenberg (2006) und Savolainen & Kari (2006) hingegen nutzten das Aufzeichnen von Bildschirmaktivitäten ihrer Teilnehmenden während der Bearbeitung von Suchaufgaben (mittels *screen capture software)*, das ebenfalls zu den Verfahren der Beobachtung zählt und insbesondere eine sinnvolle Ergänzung zur Auswertung von Logdateien darstellt (Kelly, 2009, S. 86 ff.). Eine derartige Kombination aus quantitativen und qualitativen Verfahren erfolgte in allen Studien zu Relevanzkriterien, die unter anderem Daten aus Logfiles auswerteten, bis auf die Studie von Vakkari & Hakala (2000). Das Ergebnis solcher Bildschirmaufzeichnungen sind Videodateien, die von den Forschenden auch nachträglich oder wiederholt angesehen werden können.

Von der Art der Datenerhebung unabhängig werden in den beiden folgenden Abschnitten die zwei in den 47 Studien verwendeten Forschungsdesigns erläutert. Der Fokus wird dabei der Argumentation zu Beginn dieses Kapitels folgend auf experimentelle Untersuchungsdesigns gelegt, auch wenn den wenigsten Studien zu Relevanzkriterien tatsächlich ein experimentelles Design zugrunde liegt.

2.2.3 Explorative Untersuchungsdesigns

Die Mehrheit (n = 41) der analysierten empirischen Studien zu Relevanzkriterien wurde als explorative Untersuchung durchgeführt (Tabelle 2.8). Unter ihnen sind

zudem Studien, die mitunter als Experiment vorgestellt werden, sich nach eingehender Prüfung jedoch als explorative Studie herausstellten. Auf diese Tatsache wird in Abschnitt 2.2.4 ausführlich eingegangen.

Explorative Studien dienen dazu, einen bis dato wenig erforschten Gegenstand zu erkunden und diesen zu beschreiben, und um überhaupt Hypothesen aufstellen oder Theorien bilden zu können. Zu diesem Zweck werden offene Forschungsfragen formuliert, die häufig mithilfe qualitativer Verfahren der Datenerhebung bearbeitet werden. Liegen erste Theorien und Hypothesen über den zu erforschenden Gegenstand vor, können diese in explanativen Studien, zu denen Studien mit experimentellen Designs zählen, geprüft werden. (Döring & Bortz, 2016, S. 192 ff.)

Die Voraussetzung für die Durchführung experimenteller Studien ist demnach das Vorhandensein von Erkenntnissen aus explorativen Studien. Bei den bisherigen Studien zur Erforschung von Relevanzkriterien handelt es sich überwiegend um explorative Untersuchungen. Betrachtet man die Forschungsfragen dieser Studien, wird schnell deutlich, dass sie als offene Fragen formuliert sind, dabei jedoch eine große Ähnlichkeit untereinander aufweisen, zum Beispiel:

- „What criteria do users mention when they evaluate the results of information searches in a multimedia environment?“ (Schamber, 1991);
- „What criteria allow users to determine whether connections or lack of connections exist between the information within documents and the users' information need situations?“ (Barry, 1994);
- „What are the criteria applied in judging the value of retrieved documents?“ (Wang, 1994);
- „What relevance criteria do children use to evaluate information found when searching electronic resources for a school project?“ (Hirsh, 1999);
- „How do participants judge relevance in a virtual library environment?“ (Fitzgerald & Galloway, 2001);
- „What criteria do students use when making relevance judgments about sources?“ (Twait, 2005);
- „Which criteria do students use for evaluating search results, sources, and information on a website?“ (Walraven et al., 2009);
- „What criteria do users apply in evaluating an individual document?“ (I. Xie et al., 2010);

Tabelle 2.8 Studien mit explorativen Untersuchungsdesigns

ID	Quelle	Anzahl Testpersonen	Kriterien gewichtet?	Kontext
1	Halpern & Nilan (1988)	42	Nein	ELIS
2	Nilan, Peek, & Snyder (1988)	54	Nein	ELIS
4	Schamber (1991)	30	Nein	ARB
5	Cool, Belkin, Kantor, & Frieder (1993) Studie 1	300	N. a.	AKAD
6	Cool, Belkin, Kantor, & Frieder (1993) Studie 2	11	N. a.	AKAD
7	Park (1993)	10	Nein	AKAD
8	Thomas (1993)	9	Nein	N. a.
9	Barry (1994)	18	Nein	AKAD
10	Bruce (1994)	6	Ja	AKAD
11	Howard (1994)	8	N. a.	AKAD
12	Wang (1994), Wang & Soergel (1998)	25	Nein	AKAD
13	Wang & White (1995, 1999)	15	Nein	AKAD
14	Fidel & Crandall (1997)	98	Nein	ARB
15	Markkula & Sormunen (1998)	3	Ja	ARB
16	Spink, Greisdorf, & Bateman (1998)	55	Nein	AKAD
17	Hirsh (1999)	10	Nein	SCHUL
18	Schamber & Bateman (1999); Bateman (1998, 1999)	350	Ja	N. a.
19	Vakkari & Hakala (2000)	11	Nein	AKAD
20	Fitzgerald & Galloway (2001)	10	Nein	AKAD

(Fortsetzung)

Tabelle 2.8 (Fortsetzung)

ID	Quelle	Anzahl Testpersonen	Kriterien gewichtet?	Kontext
21	Tang (1999), Tang & Solomon (2001)	99	Ja	AKAD
23	Rieh (2002)	15	Nein	AKAD/ELIS
24	Maglaughlin & Sonnenwald (2002)	12	Nein	AKAD
25	Tombros, Ruthven, & Jose (2003, 2005)	24	Nein	ELIS
26	Hung, Zoeller, & Lyon (2005)	10	Nein	ARB
27	Toms, O'Brien, Kopak, & Freund (2005)	48	Nein	AKAD/ELIS
28	Twait (2005)	13	Nein	AKAD
29	Crystal & Greenberg (2006)	12	Nein	ELIS
30	Savolainen & Kari (2006)	18	Nein	ELIS
31	Westman & Oittinen (2006)	8	Ja	ARB
33	Papaeconomou, Zijlema, & Ingwersen (2008)	15	Nein	ELIS
34	Taylor, Zhang, & Amadio (2009)	39	Nein	AKAD
35	Walraven, Brand-Gruwel, & Boshuizen (2009)	23	Nein	SCHUL
36	Beresi, et. al. (2010); Beresi (2011)	21	Nein	AKAD
37	Chu (2011)	9	Ja	AKAD
38	Balatsoukas & Ruthven (2012)	24	Nein	N. a.
40	Taylor (2012a, 2013) Studie 1	82	Nein	AKAD

(Fortsetzung)

Tabelle 2.8 (Fortsetzung)

ID	Quelle	Anzahl Testpersonen	Kriterien gewichtet?	Kontext
41	Taylor (2012b) Studie 2	64	Nein	AKAD
43	Sedghi, Sanderson, & Clough (2013)	29	Nein	ARB
44	Xie, Benoit, & Zhang (2010); Xie & Benoit (2013)	31	Nein	ARB/ELIS
45	Watson (2014)	37	Nein	SCHUL
46	Zhang (2014)	30	Nein	ELIS

N = 41

Zwar wurden in diesen Studien Kriterien in verschiedenen Kontexten untersucht, jedoch ist die Arbeit von Rieh (2002) zu den Kriterien Glaubwürdigkeit und kognitive Autorität bei der Websuche (vgl. Abschnitt 2.1.2.1) die einzige der 41 explorativen Studien, die auf die Erforschung ausgewählter Kriterien abzielte im Gegensatz zu den anderen, die offen nach allgemeinen Kriterien fragen.

Die Ergebnisse der explorativen Studien tragen zu einem besseren Verständnis der Kriterien bei der Relevanzbewertung bzw. bei der Dokumenten- bzw. Quellenauswahl bei (vgl. Abschnitt 2.1). Allerdings ist die Gesamtzahl der Kriterien und die der beeinflussenden Aspekte (Faktoren) verhältnismäßig groß, wie Xu & Chen (2006) kritisieren, und wie sich insbesondere an der Übersicht der Faktoren von Schamber (1994) (siehe Tabelle 2.1 auf S. 30) zeigt. Die Tatsache, dass bisherige Studien eher explorativ vorgehen und nicht hypothesenprüfend, nennen Xu & Chen (2006) als weiteren Kritikpunkt.

Die Besonderheiten solcher explanativen bzw. experimentellen Untersuchungen werden im nachfolgenden Abschnitt aufgezeigt; auf die Erläuterungen zur Entwicklung eines experimentellen Designs wird jedoch zunächst verzichtet, da diese im Zusammenhang mit dem in Kapitel 4 beschriebenen Online-Experiment umfassend und detailliert erfolgen.

2.2.4 Experimentelle Untersuchungsdesigns

Mithilfe von explanativen Studien werden Hypothesen über Effekte geprüft. Dabei steht die Herstellung eines Nachweises über den Zusammenhang zwischen einer Ursache und einer beobachtbaren (messbaren) Wirkung im Vordergrund.

Kausale Schlussfolgerungen über den Zusammenhang zwischen Ursache und Wirkung sind nur mithilfe von Experimenten möglich. (Döring & Bortz, 2016, S. 192)

Um kausale Schlüsse ziehen zu können, müssen drei Voraussetzungen erfüllt sein:

(a) zwei Variablen kovariieren, d. h. zwischen ihnen besteht ein nachweislicher Zusammenhang;
(b) die als Ursache vermutete Variable tritt zeitlich vor derjenigen Variablen auf, deren Wirkung beobachtet wird, denn zeitliche Präzedenz ist unerlässlich für die Bestimmung der Kausalrichtung zwischen zwei kovariierenden Variablen;
(c) alternative Erklärungen für die beobachtete Wirkung können ausgeschlossen werden (Sedlmeier & Renkewitz, 2018, S. 133 ff.).

Diese Voraussetzungen lassen sich nur durch die korrekte Planung und Durchführung eines echten Experiments erreichen. Dabei sind *Manipulation* und *Kontrolle* die beiden Wesensmerkmale eines Experiments, ohne die eine empirische Studie kein echtes Experiment ist. Echte Experimente gelten sowohl in der sozialwissenschaftlichen Grundlagen- als auch in der Anwendungsforschung als Goldstandard (Döring & Bortz, 2016, S. 102).

Bei einem Experiment wird der Effekt eines Stimulus (die unabhängige Variable, die als ursächlich vermutet wird) auf ein bestimmtes Ergebnis (die abhängige Variable) untersucht, indem die unabhängige Variable in irgendeiner Weise manipuliert wird (z. B. der Rankingalgorithmus A und B bei einem IR-System). Um alternative Gründe oder Bedingungen, die zu der Wirkung (z. B. Retrievaleffektivität) geführt haben, ausschließen zu können, müssen mögliche Störvariablen kontrolliert werden (Sedlmeier & Renkewitz, 2018, S. 124–127). Personengebundene Störvariablen, wie Alter und Einkommen, werden beispielsweise durch die zufällige Verteilung (Randomisierung) der Testpersonen auf die verschiedenen, mindestens zwei Versuchsgruppen ausbalanciert (Sedlmeier & Renkewitz, 2018, S. 134 ff.).

Der Vorteil von experimentellen Designs kann konkret am Beispiel der in Abschnitt 2.1.2.1 vorgestellten Studie zur Glaubwürdigkeit und kognitiven Autorität bei der Websuche von Rieh (2002) demonstriert werden: Die explorative Studie hätte von einem experimentellen Design profitiert, denn die Aufgabenbeschreibungen enthielten Formulierungen über die gewünschten relevanten Suchergebnisse entsprechend der operationalisierten Definition von Qualität und Autorität, wie „good papers“, „useful information“, „credible information“, „best price“. Diese Unterschiedlichkeit der Benennungen könnte die Testpersonen auf unerwünschte Weise beeinflusst und die Ergebnisse verfälscht haben, d. h. den

beobachteten Effekt alternativ erklären, wodurch eine der drei Voraussetzungen für Kausalität nicht gegeben wären. Mit einem experimentellen Design hätten beispielsweise die Adjektive jeweils als unabhängige Variablen für alle vier Aufgaben manipuliert werden können, indem eine Gruppe von Untersuchungspersonen in Aufgabe 1 die Formulierung *good papers*, eine zweite Gruppe *useful papers*, eine dritte Gruppe *credible papers* und eine vierte Gruppe *best papers* zu sehen bekommen hätte, wobei die Personen zu den einzelnen Gruppen randomisiert zugewiesen worden wären. Für die weiteren Aufgaben hätte man ebenso verfahren und den Teilnehmenden diese zusätzlich in randomisierter Reihenfolge anzeigen können.

Experimentelle Studien erlangen inzwischen auch allgemein in der Bibliotheks- und Informationswissenschaft zunehmend an Bedeutung (Connaway & Radford, 2017, S. 157), sie gelten traditionell als Hauptmethode bei der Evaluierung von Information Retrieval-Systemen (Kelly, 2009, S. 27). Das Experiment zu Relevanzbewertungen von Cuadra & Katter (1967a) gilt als klassisches Beispiel für ein echtes Experiment der Relevanzforschung: 140 Testpersonen, die 14 Versuchsgruppen (14 unterschiedliche Experimentalbedingungen) zufällig zugewiesen wurden, bearbeiteten Suchaufgaben mit manipulierten Beschreibungen von Informationsbedürfnissen und Nutzungskontexten. Die Ergebnisse zeigen, dass Relevanzbewertungen leicht zu Artefakten der jeweiligen Testinstruktionen und Bedingungen werden können und sie somit nicht in absoluten Zahlen miteinander zu vergleichen sind (Cuadra & Katter, 1967a, S. 302).

Allerdings werden im klassischen Information Retrieval, also auch im Zusammenhang mit TREC, die Begriffe Experiment[33] und Evaluation oft synonym verwendet (Kelly, 2009, S. 26), wodurch zum Teil auch Untersuchungen als Experimente bezeichnet werden, die die Bedingungen an ein Experiment nicht erfüllen. Dabei ist die Kontrolle von möglichen Störvariablen bei der Evaluierung von IIR-Systemen besonders wichtig zur Vermeidung von Positionseffekten: Beispielsweise sollte die Reihenfolge der zu bearbeitenden Suchaufgaben randomisiert werden, um mögliche Lerneffekte oder Effekte durch Ermüdung der Testpersonen zu vermeiden (Clemmensen & Borlund, 2016).

Kelly & Cresenzi bieten als Erklärung für die synonyme Verwendeung der Begriffe Experiment und Evaluation an: „IR researchers often lack formal training in the behavioral sciences and have a difficult time understanding and incorporating this perspective into their IIR experiments“ (2016, S. 1207); zugleich

[33] Während Retrieval-Experimente unter kontrollierten Bedingungen stattfinden, werden in Retrieval-Tests IR-Systeme ohne die Kontrolle von Variablen durch die Testleitung evaluiert (Tague-Sutcliffe, 1992, S. 469).

verweisen sie auf die Notwendigkeit von Experimenten: „[This perspective] is critical if we want more valid and reliable evaluations of interactive IR systems and more basic knowledge about how people interact with IR systems“ (2016, S. 1207).

Nicht alle als Experiment bezeichneten Untersuchungen erfüllen auch tatsächlich alle Anforderungen an ein echtes Experiment (auch: klassisches Experiment). Oftmals handelt es sich um Quasi-Experimente, bei denen die Reihenfolge der Testpersonen oder der zu bearbeitenden Aufgaben nicht randomisiert wurde. Dadurch sind nicht alle potenziellen Störvariablen vollständig kontrolliert, wodurch die interne Validität der Forschung gefährdet ist (Sedlmeier & Renkewitz, 2018, S. 176 ff.). Explanative Studien, denen es neben der Randomisierung auch an der Manipulation von Variablen fehlt, die aber mindestens zwei Gruppen miteinander vergleichen, werden als natürliches Experiment bzw. nicht-experimentelle Studie bezeichnet (Döring & Bortz, 2016, S. 201). Über die verschiedenen Typen eines Experiments und deren Abgrenzung zueinander gibt Tabelle 2.9 einen Überblick.

Auch in früheren Literaturstudien bzw. in den Veröffentlichungen der Studien zu Relevanzkriterien werden manche Studien als Experiment bezeichnet, obwohl es ihren Untersuchungsdesigns teilweise an Manipulation und/oder Kontrolle mangelt. Die korrekte Identifizierung der Studien, die die Anforderungen an ein (echtes) Experiment aus sozialwissenschaftlicher Sicht erfüllen, stellte die größte Herausforderung bei der Sichtung und Auswertung der ausgewählten Studien zu Relevanzkriterien dar und erforderte eine intensive Auseinandersetzung mit den in den Publikationen berichteten Methoden.

Tabelle 2.9 Typen des Experiments

	Gruppierung	Manipulation	Randomisierung
Echtes Experiment	✓	✓	✓
Quasi-Experiment	✓	✓	
Nicht-experimentelle Studie / Natürliches Experiment	✓		

Quelle: Modifiziert nach Berger & Wolbring (2015, S. 45)

Tabelle 2.10 Studien zu Relevanzkriterien mit einem experimentellen Design

ID	Quelle	Anzahl Testpersonen	Kriterien gewichtet?	Kontext
3	Regazzi (1988)	32	Ja	AKAD
22	Choi & Rasmussen (2002)	38	Ja	AKAD
32	Xu & Chen (2006)*	132	Nein	ELIS
39	De Sabbata & Reichenbacher (2012)	242	Ja	ELIS
42	Kim, Kazai, & Zitouni (2013)	28	Nein	ELIS
47	Hamid, Thom, & Iskandar (2016)	48	Ja	ARB

* Korrelationsstudie; n = 1
N = 6

Von den 47 analysierten Studien weisen nur sechs tatsächlich ein experimentelles Design auf (Tabelle 2.10), wobei eine von diesen eine Korrelationsstudie darstellt. Korrelationsstudien beschränken sich auf Erklärungen über die Art und Intensität des Kovariierens zweier Variablen und erlauben keine kausalen Schlussfolgerungen (Döring & Bortz, 2016, S. 677). Die Korrelationsstudie (auch *confirmatory study*) von Xu & Chen (2006) prüft Hypothesen zu fünf ausgewählten Kriterien (*scope, novelty, reliability, topicality, understandibility*), die in Form von Fragebogenitems den Untersuchungspersonen zur Beurteilung vorgelegt wurden. Ein Item für das Kriterium oder Konstrukt *novelty* bestand beispielsweise aus der Aussage „In this document, the amount of new information to me is ____“, zu der die Probandin/der Proband auf einer 7-stufigen Skala zwischen *Small* (1) und *Substantial* (7) ihre/seine Antwort kennzeichnete (Xu & Chen, 2006, S. 972). Hieran wird der bedeutende Unterschied im Vergleich zu explorativen Untersuchungen deutlich: Explanative und experimentelle Studien erfordern eine Variable (hier: *novelty*), die es zu operationalisieren gilt (hier: *amount of new information*), wofür wiederum eine eindeutige Definition des betreffenden Konstrukts vorhanden sein muss.

In Anbetracht der Schwierigkeiten bei der Benennung und Abgrenzung einzelner Relevanzkriterien untereinander und im Zusammenhang mit Faktoren, die die Relevanzbewertung beeinflussen (vgl. Abschnitt 2.1.5), wird hier bereits eine weitere Forschungslücke hinsichtlich der Operationalisierbarkeit der zu untersuchenden Variablen sichtbar.

Von den fünf weiteren experimentellen Studien wurden in zwei Studien den Teilnehmenden Suchergebnisse in einem akademischen Kontext zur Bewertung vorgelegt. Regazzi (1988) ließ 32 Forschende und Studierenden aus den Bereichen Biomedizin und Sozialwissenschaften insgesamt 16 Dokumente zum Thema Alkohol unter Verwendung einer 3-Punkte-Skala bewerten. Das Hauptziel der Studie bestand darin, anhand des Vergleichs von Bewertungen unterschiedlicher Gruppen von Juroren zu jeweils 4 Aufgaben mit jeweils 4 verschiedenen Dokumenten Rückschlüsse auf geeignete Kennzahlen zur Performanz bibliographischer Informationssysteme zu ziehen. So bewertete eine Gruppe die Dokumente nach deren Relevanz, die anderer Gruppe bewertete deren Nützlichkeit. Zusätzlich sollten die Teilnehmenden Aussagen über die Wichtigkeit einzelner Elemente der zu bewertenden Dokumente treffen; diese Elemente waren allerdings nicht manipuliert, sodass diese streng genommen keinen zu untersuchenden Faktor im Experiment darstellten. Der Verzicht auf diesen Faktor ist aufgrund des ohnehin komplexen Designs der Studie nachvollziehbar; jedoch muss dies beim Bewerten der Studienergebnisse berücksichtigt werden. So sind kausale Schlussfolgerungen zwischen den Elementen der Dokumente auf die Bewertungen nicht möglich.

Choi & Rasmussen (2002) ließen 18 Fakultätsangehörige und 20 Studierende amerikanischer Geschichte Bilder und dazugehöriger Metadaten in Form von Textinformationen eines Fotoarchivs bewerten. Die Untersuchungspersonen wurden zufällig zwei Gruppen zugeteilt: In Gruppe A wurden den Teilnehmenden zuerst die Bilder gezeigt, welche mithilfe eines Fragebogens beurteilt wurden und im Anschluss wurden die Metadaten zu einer weiteren Bewertung offengelegt. Die Teilnehmenden in Gruppe B sahen zuerst die Metadaten, gaben ihre Bewertung ab, bevor sie das Bild sahen, welches sie ebenfalls bewerteten. Mit diesem Vorgehen wurde der Effekt des Hinzufügens bestimmter Bilder oder Textinformationen auf die Bewertung untersucht, indem geprüft wurde, ob sich die Bewertungen nach dem Hinzufügen der neuen Informationen verändert hatten oder nicht. Die Bewertungen wurden anhand einer 7-stufigen Skala erhoben.

Manipuliert wurde in dem Experiment somit die Reihenfolge der den Teilnehmenden präsentierten Informationen (Metadaten und Bild), die Metadaten an sich wurden nicht variiert. Daher lassen sich auch hier keine kausalen Schlüsse über den Effekt der (jeweiligen) Metadaten auf die Bewertung folgern. Ungeachtet dessen zeigten die Ergebnisse keine statistisch signifikanten Unterschiede zwischen den Bewertungen vor und nach dem Hinzufügen der Bilder bzw. der Textinformationen. Allerdings verdeutlichen die Ergebnisse den besonderen Stellenwert eines Bild-Surrogats für die Relevanzbewertung: Den Teilnehmenden in Gruppe A gelang es sehr häufig nicht, eine Bewertung ausschließlich anhand des Bildes vorzunehmen, ohne dessen Metadaten zu berücksichtigen. Im Gegensatz

zu anderen Studien zu Relevanzkriterien, in denen rein textbasierte Dokumente bewertet wurden, war den Teilnehmenden die Person, die als Urheber der jeweiligen Fotographie gilt, nicht wichtig für die Bewertung (Choi & Rasmussen, 2002, S. 711).

Mit der Bewertung von Bildern beschäftigten sich auch Hamid et al. (2016). In ihrer experimentellen Studie baten sie 48 Studierende unterschiedlicher Fachdisziplinen um eine Bewertung von Bildern der Google-Bildersuche mit einem realen Bezug zu ihrer Arbeit. Daten zum Such- und Bewertungsverhalten wurden mithilfe eines Fragebogens sowie einer Screen-Capture-Software (technikvermittelte Beobachtung). Die Testpersonen bearbeiteten jeweils insgesamt vier Aufgaben aus vier verschiedenen Aufgabentypen, wobei jede Testperson einem Aufgabentyp randomisiert zugewiesen wurde, die Person allerdings eine Aufgabe innerhalb dieses Typs frei wählen konnte. Anschließend wurde den Testpersonen eine Liste der folgenden 10 Relevanzkriterien (in Form von Selbstaussagen) vorgelegt, die sie bezüglich ihrer Bildbewertung nach deren Wichtigkeit beurteilen sollten (Hamid et al., 2016, S. 6):

- Topicality (The images I selected were relevant to the search topic.)
- Accuracy (The images I selected were an accurate representation of what I was looking for on the search topic.)
- Suggestiveness (The images I selected gave me new ideas or new insights about the search topic.)
- Appeal of information (The images I selected were interesting in regards [sic] to the search topic.)
- Completeness (The images I selected contained the kinds of details I could use to clarify important aspects of the search topic.)
- Technical attributes (The images I selected had technical attributes (such as colour, perspective or angle) that were important to me for this search topic.)
- Emotion (The images I selected evoked an emotional response in me regarding the search topic.)
- Textual information (The images I selected had useful text descriptions on the search topic.)
- Consequence (The images I selected contained consequences or implications of the search topic.)
- Composition (The images I selected have a strong visual impact regarding the search topic.)

Somit sind auch bei dieser experimentellen Studie keine kausalen Schlussfolgerungen in Hinblick auf den Effekt von (ausgewählten) Relevanzkriterien möglich,

da diese den Testteilnehmenden vorgelegt und nicht variiert wurden. Die subjektive Beurteilung der Kriterien nach deren Gewichtung war demzufolge auch bei dieser Studie nicht Gegenstand des eigentlichen experimentellen Designs.

2.2.5 Zusammenfassung

Die Analyse der Methoden von 47 Studien zu Relevanzkriterien ergab, dass die bisherige Erforschung von Relevanzkriterien überwiegend auf explorativen Studien beruht, die offene Forschungsfragen anhand von Daten, die mit Verfahren der Befragung erhoben wurden, bearbeiteten. Zu den verwendeten Verfahren zählen häufig die Methode des lauten Denkens und die anschließende Analyse der dabei entstandenen Thinking-aloud-Protokolle, seltener Tagebucheinträge und Fokusgruppeninterviews. Nicht in allen Studien wurden explizite Relevanzbewertungen erfasst, jedoch sind auch implizit ausgedrückte Bewertungen, wie bei der Methode des lauten Denkens, analysierbare Relevanzbewertungen.

Während alle 47 Studien mindestens eine Form der Befragung durchführten, wurden in 17 Studien zusätzlich Verfahren der Beobachtung zur Datenerhebung genutzt. Diese sind Bildschirmaufzeichnungen von Interaktionen während des Suchprozesses, Eye-Tracking-Studien zur Analyse von Blickbewegungen sowie die Auswertung von Logfiles bzw. Klicks (technikvermittelte Beobachtung), wie sie häufig in der Interactive Information Retrieval-Forschung zum Einsatz kommen.

Jüngere Studien bauen oft auf den Erkenntnissen der explorativen Studien der 1990er Jahre auf, welche somit eine wichtige Grundlage für nachfolgende Studien und auch weiterhin für zukünftige Forschungsvorhaben zu Relevanzkriterien darstellen. Der Vorteil von explorativen Studien liegt in der Offenheit der Forschungsfragen, die erforderlich sind, wenn keine Hypothesen oder bestehenden Theorien über den zu untersuchenden Gegenstand existieren oder diesen unzureichend beschreiben. Aufgrund der Vielzahl inzwischen erkannter Kriterien erscheint es ein Versäumnis, nicht hypothesenprüfend vorzugehen.

Als Goldstandard hypothesenprüfender, explanativer Studien gilt das Experiment, wie es in den Sozialwissenschaften oder der Psychologie häufig zum Einsatz kommt. Ein Experiment ist die beste Methode, um kausale Schlussfolgerungen über Ursache-Wirkungszusammenhänge ziehen zu können. Bezogen auf den Forschungsgegenstand der Relevanzkriterien sind kausale Schlüsse über den Zusammenhang zwischen den bei der Bewertung von Informationsobjekten verwendeten subjektiven Kriterien und den beobachteten Unterschieden bei den Relevanzbewertungen nur bei Studien mit einem echten experimentellen Design

möglich. Insgesamt weisen lediglich sechs der 47 Studien ein experimentelles Forschungsdesign auf. Keine dieser Studien erlaubt Rückschlüsse über die Wirkung der Relevanzkriterien auf die Relevanzbewertung aufgrund einer fehlenden Manipulation von Kriterien. In keiner der sechs Studien wurde explizit der Einfluss mehrerer Kriterien oder eines bestimmten Kriteriums als unabhängige Variable(n) auf die Relevanzbewertung als abhängige Variable untersucht.

2.3 Fazit und Forschungsfragen

In diesem Abschnitt werden die mithilfe der Betrachtung des Forschungsstands identifizierten Forschungslücken dargelegt. Auf eine erneute Zusammenfassung des Forschungsstands aus der inhaltlichen und methodischen Perspektive wird an dieser Stelle verzichtet. Direkt anknüpfend an den vorhergehenden Abschnitt wird zunächst die Forschungslücke aus der methodischen Sichtweise auf den Forschungsstand beleuchtet:

In Hinblick auf die Verwendung experimenteller Untersuchungsdesigns in Studien zu Relevanzkriterien wurde aufgezeigt, dass Studien zu Relevanzkriterien überwiegend einen explorativen Ansatz verfolgen; nur in wenigen Studien wurde ein Experiment durchgeführt. Ein experimentelles Design ist jedoch die einzige Möglichkeit, kausale Schlussfolgerungen über Ursache-Wirkungszusammenhänge wie den Einfluss von Relevanzkriterien auf die Bewertung von Informationsobjekten ableiten zu können. Mithilfe experimenteller Untersuchungen werden Hypothesen geprüft, während explorative Studien offene Forschungsfragen bearbeiten. Inzwischen sind auf der Basis der Erkenntnisse der zahlreichen explorativen Studien diverse Kriterien aufgedeckt worden, sodass sich Hypothesen über die konkrete Verwendung ausgewählter Relevanzkriterien aufstellen lassen. Von den bisherigen experimentellen Studien zu Relevanzkriterien beziehen sich lediglich zwei Studien auf die Relevanzbewertung im akademischen Kontext; zusätzliche Daten wie Popularitätsdaten waren nicht Bestandteil der von den Jurorinnen und Juroren bewerteten Suchergebnisse. Zudem untersuchten diese Studien nicht gezielt den Einfluss bestimmter Kriterien auf die Relevanzbewertung, da diese Kriterien nicht als unabhängige Variablen manipuliert wurden. Diese Forschungslücke soll mit der folgenden Forschungsfrage bearbeitet werden:

F1 Wie können Nutzerkriterien bei der Relevanzbewertung anhand eines experimentellen Untersuchungsdesigns erforscht werden?

Die Betrachtung des Forschungsstands zu Relevanzkriterien aus inhaltlicher Perspektive zeigt zwei konkrete Forschungslücken auf, von denen nachfolgende in direktem Zusammenhang mit der Bearbeitung der Forschungsfrage F1 steht:

Es gibt keine definitorische und konzeptuelle Abgrenzung der Begriffe Merkmale (clues, cues), Kriterien (criteria) und Faktoren (factors). Diese Begriffe werden in der Literatur zur Erforschung von Kriterien und Einflussfaktoren unterschiedlich verwendet und oft nicht eindeutig definiert. Oftmals taucht in der Literatur zur Erläuterung des Begriffs Kriterium die Formulierung „Gründe für die Bewertung" (*underlying reasons behind relevance judgments*) auf; allerdings ist diese Bezeichnung sehr breit, denn Gründe können in Hinblick auf alle drei Begriffe Merkmal, Kriterium und Faktor genannt werden. Zudem lässt sich anhand der Begriffe, die als Merkmale, Kriterien oder Faktoren ausgewiesen werden, aufgrund der Verwendung von Synonymen und Homonymen keine genaue Zuweisung zu den drei genannten Gruppen vornehmen. Diese fehlenden Definitionen und die ungenaue Abgrenzung der Begrifflichkeiten erschweren es, Ergebnisse aus bisherigen Studien zu Relevanzkriterien einzuordnen und zu analysieren. Für systematische Literaturschauen oder Metaanalysen ist es daher nicht möglich, ein klares Bild über die Zusammenhänge der die Relevanzbewertung beeinflussenden Aspekte zu erhalten, ohne dass die Gefahr besteht, aufgrund von Fehlinterpretationen durch unklare Terminologie falsche Schlussfolgerungen zu ziehen. Demzufolge ist die Klärung der Begriffe Kriterien, Faktoren und Merkmale sowie deren eindeutige Unterscheidung die Voraussetzung für die Durchführung einer hypothesenprüfenden Studie, für die die untersuchten Variablen operationalisiert werden müssen. Ohne eine eindeutige Definition des Begriffs Kriterium im Kontext der Relevanzbewertung, die mit der konzeptuellen Abgrenzung zu den anderen Begriffen Faktor und Merkmal einhergeht, ist es nicht operationalisierbar und somit nicht messbar (Döring & Bortz, 2016, S. 224 ff.). Diese Forschungslücke soll mit der Beantwortung der folgenden Unterforschungsfragen geschlossen werden:

F1a Wie lassen sich Merkmale, Kriterien und Faktoren als Einflüsse im Prozess der Relevanzbewertung für die Entwicklung eines experimentellen Untersuchungsdesigns definitorisch und konzeptuell voneinander abgrenzen?

F1b Wie können Kriterien bei der Relevanzbewertung von Suchergebnissen für eine experimentelle Studie operationalisiert werden?

Die Beantwortung der beiden Unterforschungsfragen F1a und F1b ist die Voraussetzung für die Bearbeitung und Beantwortung der Forschungsfrage F1,

die mit dem Ziel einhergeht, ein nachnutzbares methodisches Framework zur experimentellen Erforschung von Relevanzkriterien zu entwickeln.

Zusätzlich zu der Beantwortung der Forschungsfrage F1 zielt das Experiment auf der inhaltlichen Ebene auf die Beantwortung zweier weiterer Forschungsfragen ab, die sich aus der zweiten Forschungslücke bei der inhaltlichen Betrachtung des Forschungsstands ergeben:

In den bisherigen Studien, in denen die Erkenntnisse über Relevanzkriterien auf Relevanzbewertungen, die von den Studienteilnehmenden auf Basis von Surrogaten in akademischen Suchsystemen vorgenommen wurden, beruhen, enthielten die Surrogate neben den erschließungstypischen Metadaten keine zusätzlichen Informationen wie Popularitätsdaten, die die Bewertung ebenfalls beeinflussen können. Heutzutage sind in modernen akademischen Suchsystemen Popularitätsdaten wie die Anzahl von Downloads und die Anzahl von Zitationen eines Werks in die Suchergebnisdarstellung als zusätzliche Metadaten integriert. Diese Informationen können Hinweise beispielsweise über die Autorität einer Autorin oder eines Autors liefern und damit einen Indikator für die zu erwartende Qualität des Werks darstellen. Unklar ist, in welcher Weise solche Popularitätsdaten die Relevanzbewertung in akademischen Suchsystemen beeinflussen. Diese Forschungslücke führt zu der folgenden Forschungsfrage:

F2 Welchen Einfluss haben Popularitätsdaten auf die Bewertung der Relevanz von Suchergebnissen in akademischen Suchsystemen?

Wenn ein Einfluss von Popularitätsdaten als sichtbarer Bestandteil der Suchergebnisdarstellung auf die Relevanzbewertung experimentell nachgewiesen wird, kann dieser positiv oder negativ sein. Vermutet wird ein allgemeiner positiver Einfluss von Popularitätsdaten, der sich in einer höheren Relevanzbewertung niederschlägt. Diese Vermutung stützt sich insbesondere auf die Ergebnisse der Arbeiten von Rieh (2002) und Wang (1994) hinsichtlich der Bedeutung von zusätzlichen Informationen in einem Suchergebnis und den besonderen Stellenwert von Glaubwürdigkeit und Autorität auf der Basis von Informationen bzw. Wissen über die Autorin / den Autor eines Werkes. Die Annahme eines positiven Einflusses von Popularitätsdaten auf die Relevanzbewertung findet Ausdruck in den inhaltlichen Hypothesen, die im Rahmen der Entwicklung des experimentellen Designs aufgestellt werden.

Für ein differenzierteres Bild werden unterschiedliche Arten von Popularitätsdaten, also mehr als eine unabhängige Variable, untersucht. Die letzte Forschungsfrage zielt daher auf die Gewichtung dieser verschiedenen Popularitätsdaten ab:

F3 Welche Popularitätsdaten beeinflussen die Relevanzbewertung in welchem Maße?

Da bisherige Studien zeigen, dass diverse Relevanzkriterien im Prozess der Relevanzbewertung eine Rolle spielen, stellt sich die Frage, wie diese Kriterien gewichtet werden. Unabhängig von dem Kriterium der thematischen Relevanz als Basis für die Bewertung ist unklar, welchen Stellenwert das Kriterium Popularität einnimmt. Diese Lücke lässt sich mit Beantwortung dieser Forschungsfrage zwar nicht schließen, bildet jedoch einen wichtigen Baustein zur Klärung der Frage, wie die unterschiedlichen Arten von Popularitätsdaten, anhand derer sich das Kriterium der Popularität ableiten lässt, zusammenwirken. Es besteht die Annahme, dass nicht alle Popularitätsdaten den gleichen, positiv vermuteten Effekt auf die Relevanzbewertung bewirken.

Eine systematische Gegenüberstellung der identifizierten Forschungslücken mit den Forschungsfragen, die jeweils auf die Schließung der Forschungslücke abzielen, bietet Tabelle 2.11.

Tabelle 2.11 Gegenüberstellung der Forschungslücken und Forschungsfragen

Forschungslücke	Forschungsfrage
L1: Sehr wenigen bisherigen Studien zu Relevanzkriterien liegt ein experimentelles Untersuchungsdesign zugrunde; Kriterien wurden nicht als unabhängige Variablen variiert.	F1:Wie können Nutzerkriterien bei der Relevanzbewertung anhand eines experimentellen Untersuchungsdesigns erforscht werden?
L2: Es gibt keine definitorische und konzeptuelle Abgrenzung der Begriffe Merkmal, Kriterium und Faktor.	F1a: Wie lassen sich Merkmale, Kriterien und Faktoren als Einflüsse im Prozess der Relevanzbewertung definitorisch und konzeptuell voneinander abgrenzen?
	F1b: Wie können Kriterien bei der Relevanzbewertung von Suchergebnissen für eine experimentelle Studie operationalisiert werden?
L3: In bisherigen Studien, in denen den Teilnehmenden Surrogate zur Relevanzbewertung vorgelegt wurden, waren Popularitätsdaten nicht Bestandteil der Suchergebnispräsentation.	F2: Welchen Einfluss haben Popularitätsdaten auf die Bewertung der Relevanz von Suchergebnissen in akademischen Suchsystemen?
	F3:Welche Popularitätsdaten beeinflussen die Relevanzbewertung in welchem Maße?

Im nachfolgenden Kapitel 3 werden zunächst die Voraussetzungen zur experimentellen Erforschung von Relevanzkriterien mit der Bearbeitung der Unterforschungsfragen F1a und F1b bearbeitet, welche in Abschnitt 3.3 konkret beantwortet werden. Daran schließt sich mit Kapitel 4 der Kern der vorliegenden Arbeit an, in dem die experimentelle Studie zur Untersuchung des Einflusses von Popularitätsdaten auf die Relevanzbewertung von Suchergebnissen in akademischen Suchsystemen beschrieben wird. Dies führt zur Beantwortung der Forschungsfragen F1, F2 und F3 im anschließenden Abschnitt 5.1.

3 Voraussetzungen zur experimentellen Erforschung von Relevanzkriterien

Für die Entwicklung eines experimentellen Studiendesigns müssen die zu untersuchenden Variablen operationalisiert werden. Einer Operationalisierung muss die Definition des Forschungsgegenstands vorangehen (Döring & Bortz, 2016, S. 222 ff.). Gegenstand der vorliegenden Arbeit ist der subjektive Prozess der Relevanzbewertung, der von verschiedenen Parametern beeinflusst wird und als aufeinander folgende Anwendung von Relevanzkriterien (Beresi et al., 2010, S. 199) zu verstehen ist. Um diesen Prozess besser zu verstehen, ist das Wissen über die Bildung und Anwendung von Kriterien bei der Relevanzbewertung notwendig (vgl. Abschnitt 1.1). Relevanzkriterien und Relevanzbewertung zu definieren, erfordert eine Auseinandersetzung mit dem informationswissenschaftlichen Konzept von Relevanz. Diese Auseinandersetzung geschieht im Rahmen der Konzeptspezifikation, die im nachfolgenden Abschnitt 3.1 erfolgt.

Die Spezifikation beginnt mit der Darstellung verschiedener Relevanzformen, die die informationswissenschaftliche Relevanzforschung hervorgebracht hat; das Verständnis dieser Formen unterstützt die Definitionsfindung von Relevanz, welche der im weiteren Verlauf dieser Arbeit vorgestellten empirischen Studie den theoretischen Rahmen liefert (Abschnitt 3.1.1). Die Berücksichtigung der Multidimensionalität von Relevanz trägt ebenfalls zur Konzeptspezifikation bei (Abschnitt 3.1.2). Daran anknüpfend wird die Auffassung der Relevanzbewertung von Suchergebnissen als Beurteilungsprozess in Abgrenzung zur Auswahl von Dokumenten als Entscheidungsprozess begründet (Abschnitt 3.1.3). Die Relevanzbewertung als Prozess des Urteilens zu verorten, ebnet den Weg zu einer definitorischen Abgrenzung der Begriffe Relevanzmerkmale (*relevance clues*), Relevanzkriterien (*relevance criteria*) und Relevanzfaktoren (*relevance factors*), die als Einflüsse im Prozess der Relevanzbewertung von Surrogaten identifiziert werden (Abschnitt 3.2), insofern, dass Relevanzkriterien keine Entscheidungskriterien darstellen, sondern Kriterien für die Urteilsbildung, wobei das Urteil zu

C. Behnert, *Popularität und Relevanz in der Suche*,
https://doi.org/10.1007/978-3-658-37512-6_3

einer Entscheidung führt. Da hier ausschließlich *predictive judgments* betrachtet werden, wird auf eine Bezugnahme zu Urteilen und Entscheiden bei *evaluative judgments* verzichtet.

Das Ergebnis von Abschnitt 3.2 ist ein Modell zur subjektiven Relevanzbewertung (Abbildung 3.4), das einen Lösungsvorschlag für das Definitionsproblem auf Basis der Erkenntnisse aus der Literaturschau (vgl. Abschnitt 2.1) bietet und die Einflussparameter bei der Relevanzbewertung von Suchergebnissen in akademischen Suchsystemen systematisch in ihrem Zusammenspiel aufzeigt. Auf diese Weise stellt das Modell ein Hilfsmittel zur Operationalisierung der für die in Kapitel 4 vorgestellten Studie dar und kann für künftige Studien nachgenutzt werden. Abschließend erfolgt die Beantwortung der Forschungsfragen F1a und F1b (Abschnitt 3.3).

3.1 Spezifikation des Relevanzkonzepts

Die in diesem Abschnitt beschriebene Konzeptspezifikation knüpft an die Erläuterungen zum Relevanzkonzept in der Problemdarstellung der Einleitung an (vgl. Abschnitt 1.1). Wie bereits erläutert, ist das informationswissenschaftliche Relevanzkonzept sehr komplex und wird traditionell aus zwei verschiedenen Perspektiven betrachtet – der systemseitigen und der nutzerseitigen Sicht auf Relevanz. Die nutzerseitige, subjektive Perspektive auf Relevanz bestimmte die Betrachtung des Forschungsstands zu Relevanzkriterien (vgl. Kapitel 2) und wird im nachfolgenden Abschnitt 3.1.1 hinsichtlich der unterschiedlichen Formen weiterhin eingenommen. Das Wissen um die verschiedenen Formen oder Arten von Relevanz ist für eine entsprechende Verortung und letztendlich für die Definition von Relevanz notwendig.

3.1.1 Relevanzformen und Relevanzdefinition

Charakterisierend für die Komplexität der nutzerbasierten Sicht auf Relevanz sind die verschiedenen Begriffe, die in der Literatur als **Formen von Relevanz** bezeichnet werden. Diese verdeutlichen die Vielfalt, mit der versucht wurde, das Konzept von Relevanz wissenschaftlich zu durchdringen. Die Relevanzformen vereint neben dem dynamischen und kontextabhängigen Charakter der bedürfnisorientierte Definitionsansatz von Relevanz, der sich von dem anfrageorientierten Ansatz der systemseitigen Relevanzperspektive abgrenzt. So bezeichnet

Pertinenz (*pertinence*) die Beziehung zwischen dem menschlichen Informationsbedürfnis[1] und einem Informationsobjekt, die demzufolge nur durch die informationssuchende Person selbst beurteilt werden kann und die Dynamik des menschlichen Informationsbedürfnisses respektiert (Borlund, 2003b). Die Beurteilung von Pertinenz erfordert kognitive Ressourcen, folglich wird Pertinenz auch als **kognitive Relevanz** (*cognitive relevance*) bezeichnet (Saracevic, 1996) und betont damit die kognitiven Veränderungen im Menschen während der Informationssuche und der Bewertung der Ergebnisse. Der Begriff **situative Relevanz** (*situational relevance*) wurde von Wilson (1973) eingeführt und bezeichnet die Relevanzbeziehung zwischen einem Informationsobjekt und einem bestimmten Problem bzw. Anliegen (*concern*) in Bezug auf die subjektive Sicht eines Individuums auf dessen Situation. Das bedeutet, dass Informationsobjekte situativ relevant sind, wenn sie zumindest dabei helfen, Fragen bezüglich des Problems (*questions of concern*) zu beantworten (P. Wilson, 1973). Da sich die Sichtweise der suchenden Person in Hinblick auf ihre Situation verändern kann, ist situative Relevanz ebenfalls stark dynamisch. Eine Form der nutzerbasierten Relevanz, die im Prinzip kognitive und situative Relevanz zusammenfasst (Borlund, 2003b), ist die **psychologische Relevanz** (*psychological relevance*). Nach der Definition von Harter (1992)[2] bezeichnet diese die Beziehung zwischen einem Informationsobjekt und dem jeweiligen psychologischen Zustand, in dem sich die informationssuchende Person zu dem Zeitpunkt befindet. Da sich der mentale Zustand der Person mit jedem neu hinzukommenden, als relevant beurteilten Informationsobjekt verändert, ist auch psychologische Relevanz nicht als statistisch, sondern als dynamisch zu betrachten. Ein Informationsobjekt ist für die suchende Person in ihrem Kontext psychologisch relevant, wenn es kognitive Veränderungen hervorruft (Harter, 1992).

Neben Pertinenz bzw. kognitiver Relevanz, situativer Relevanz und psychologischer Relevanz wurde das Konzept der **Nützlichkeit** (*utility*, auch *usefulness*[3])

[1] Das Konzept des Informationsbedürfnisses (*information need*) nach R. S. Taylor (1968) brachte den Begriff Pertinenz hervor (Saracevic, 1975); im Zusammenhang mit Relevanz taucht der Begriff *pertinent* erstmals bei Königová (1971) auf.

[2] Harter (1992) greift hierfür auf die ursprünglich von Sperber und Wilson 1986 veröffentlichte Relevanztheorie zurück, welche Relevanz als Beziehung zwischen einer gegebenen Annahme und einem gegebenen Kontext im Rahmen von verbaler Kommunikation erachtet. Er überträgt diesen Ansatz auf den IR-Kontext. Zur kommunikationswissenschaftlichen Relevanztheorie von Sperber und Wilson siehe Sperber & Wilson (1995) und D. Wilson & Sperber (2004). Eine informationswissenschaftliche Relevanztheorie gibt es bislang nicht (Saracevic, 2016a).

[3] *Usefulness*, also Nützlichkeit, wurde alternativ zu Relevanz als Basiskriterium für die Evaluierung im IIR vorgeschlagen, um den gesamten interaktiven Informationssuchprozess in

eingeführt. Nach der Argumentation von Soergel (1994) setzt Nützlichkeit Pertinenz voraus: Ein Informationsobjekt ist dann nützlich, wenn es pertinent ist und zusätzlich sein Informationsgehalt zur Erweiterung des bereits vorhandenen Wissens der Person beiträgt (Soergel, 1994). Somit beinhaltet Nützlichkeit den kognitiven Aspekt von psychologischer Relevanz. Dagegen definieren Saracevic (1996) und Borlund (2003) situative Relevanz als die Wahrnehmung der Nützlichkeit (*utility*) eines Informationsobjekts in Bezug zur Situation der betreffenden Person.

Die Berücksichtigung der Situation, in der sich eine Person zum Zeitpunkt der Informationssuche befindet, ist innerhalb der bisher genannten Konzepte begrenzt auf das Informationsbedürfnis. Die Situation kann auch auf den übergeordneten Prozess der Aufgabenbearbeitung bzw. Problemlösung (*task process*) ausgeweitet werden: Reid (1999) greift Relevanz als ziel- bzw. aufgabenorientiertes Konzept auf und führt den Begriff der **Aufgabenrelevanz** (*task relevance*) ein. Nach ihrer Definition ist ein Dokument aufgabenrelevant (*task relevant*), wenn die Informationen, die es enthält, zur Erledigung der Aufgabe bzw. Lösung des Problems beitragen. Die Aufgabenrelevanz kann demnach erst zu dem Zeitpunkt bestimmt werden, wenn die informationssuchende Person die Aufgabe ausgeführt hat, also sowohl nach Abschluss des Informationssuchprozesses als Teilprozess der Aufgabenbearbeitung, als auch nach Sichtung und Studium der ausgewählten Quellen. Reid (1999) argumentiert, dass Aufgabenrelevanz die „ultimative“ Relevanz auf Nutzerseite darstellt, da sie den Kontext und das reale Informationsbedürfnis zu einem Zeitpunkt berücksichtigt, zu dem das Informationsbedürfnis befriedigt ist, nämlich dann, wenn die eigentliche Aufgabe abgeschlossen bzw. das Problem gelöst wurde. Borlund (2003) setzt Aufgabenrelevanz mit situativer Relevanz gleich, obwohl situative Relevanz die übergreifende Situation, also den *task process*, nicht explizit berücksichtigt. Den Aufgabenbezug von Relevanz stellen auch Hjørland & Christensen (2002) her. Sie definieren ähnlich wie Reid (1999) Relevanz als zielorientiert: „Something (A) is relevant to a task (T) if it increases the likelihood of accomplishing the goal (G), which is implied by T“ (Hjørland & Christensen, 2002, S. 964). Diese aufgaben- und zielorientierte Definition von Relevanz lässt allerdings den Zeitpunkt der Relevanzbewertung offen. In Hinblick auf den Prozess der Relevanzbewertung von Suchergebnissen kann die **Relevanzdefinition** in Hinblick auf die Art der Bewertung (*predictive judgment*)

Hinblick auf die Lösung des Informationsproblems und die Erreichung des dahinter liegenden Ziels der informationssuchenden Person zu berücksichtigen (Belkin, 2015; Cole u. a., 2009).

und somit bezogen auf die Art des Bewertungsgegenstands, des Surrogats, wie folgt modifiziert werden:

A surrogate (A) is relevant to a task (T) if it is perceived as increasing the likelihood of accomplishing the goal (G), which is implied by T. – Ein Surrogat (A) ist für eine Aufgabe (T) relevant, wenn es als die Wahrscheinlichkeit erhöhend wahrgenommen wird, das Ziel (G) zu erreichen, welches durch T impliziert wird. Diese Definition ist unabhängig von dem Kontext des betreffenden Suchsystems, d. h. sie gilt für akademische und nicht-akademische Suchsysteme gleichermaßen, solange diese Systeme Surrogate als Suchergebnisse präsentieren. Dass ein Surrogat für eine Aufgabe relevant sein kann, deutet zwar eine dichotome Ausprägung von Relevanz an, ist hier unter Berücksichtigung der *graded relevance*[4] allerdings als Abstufungen implizierend zu verstehen.

Abschließend soll an dieser Stelle ein weiterer Aspekt des Relevanzkonzepts erwähnt werden: Den Begriff gesellschaftliche Relevanz (*societal relevance*) erläutern Haider & Sundin (2019) auch als gesellschaftliches Interesse (S. 10) am Beispiel der Suche im Web nach Informationen zum Thema Impfen, bei dem die Einstellung des Einzelnen im Extremfall vollständig von dem gesamtgesellschaftlichen Interesse basierend auf einem wissenschaftlich fundierten Konsens abweichen kann.

Trotz oder gerade wegen dieser in Abbildung 3.1 dargestellten Vielzahl an Begrifflichkeiten und Relevanzformen sind bisherige Versuche darüber, eine allgemeingültige Definition von Relevanz zu formulieren, gescheitert (Saracevic, 2016b). Es ist davon auszugehen, dass die unterschiedlichen Perspektiven und die Komplexität des Relevanzkonzepts die Festlegung auf eine Definition, welche alle Aspekte berücksichtigt, erschweren (Borlund, 2003b; Saracevic, 2016b; Schamber et al., 1990). Für die vorliegende Arbeit wird – wie oben dargelegt – die nach Hjørland & Christensen (2002) modifizierte ziel- und aufgabenorientierte Relevanzdefinition verwendet.

[4] Das Konzept der *graded relevance* findet insbesondere Berücksichtigung im Zusammenhang mit der Erhebung expliziter Relevanzbewertungen anhand von mehrstufigen Skalen im Gegensatz zu einer Skala mit lediglich zwei Ausprägungen (relevant – nicht relevant); darauf wird in Abschnitt 4.1.2.1 näher eingegangen.

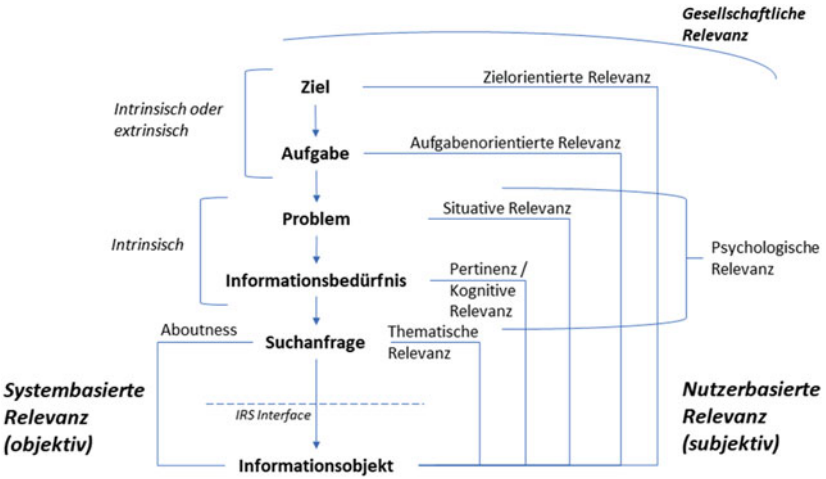

Abbildung 3.1 Übersicht über Relevanzformen und Relevanzperspektiven

3.1.2 Multidimensionalität von Relevanz

Die verschiedenen Relevanzformen und Relevanzbeziehungen kennzeichnen die Multidimensionalität des Relevanzkonzepts. Deswegen ist es von großer Bedeutung, explizit zu benennen, um welche Form von Relevanz bzw. um welche Relevanzbeziehung es sich in dem jeweiligen (Forschungs-)Kontext handelt, um Missverständnissen vorzubeugen (Mizzaro, 1997). Vor diesem Hintergrund spricht Mizzaro (1997) sogar von einer Vielzahl von Relevanzen (*relevances*) und setzt somit die Begriffe Relevanzkonzept und Relevanzbeziehung in ein synonymisches Verhältnis. In seinem Framework stellt er Relevanz als Beziehung zwischen den Entitäten zweier Dimensionen dar (siehe Abbildung 3.2): Die I. Dimension beinhaltet ein Surrogat, ein Dokument, eine Information, wie sie durch die Interaktion einer informationssuchenden Person mit einem IRS produziert werden; die II. Dimension zeigt die Nutzerseite, von der Wahrnehmung eines Informationsproblems über das Erkennen des Informationsbedürfnisses und die Formulierung der Suchanfrage, zunächst natürlichsprachlich, schließlich formalisiert als Anfrage an das IRS. Eine Relevanzbeziehung besteht beispielsweise zwischen einem Informationsbedürfnis und einem Dokument oder zwischen

einer Suchanfrage (natürlichsprachlich oder formalisiert) und einem Surrogat. In Abbildung 3.2 deuten die hellgrau gestrichelten Pfeile zwischen den jeweiligen Entitäten innerhalb der I. und II. Dimension den Prozess der Informationssuche an – von der Problemerkennung auf Nutzerseite über die Eingabe der Suchanfrage bis zum ausgewählten Dokument, welches schließlich die gesuchte Information enthält. Jede der einzelnen Entitäten und ihrer (Relevanz-)Beziehungen untereinander ist immer hinsichtlich der Thematik, der Suchaufgabe und des Kontexts zu berücksichtigen; Mizzaro (1997) fasst diese drei Aspekte als III. Dimension zusammen. Die IV. Dimension ist die Zeit, die zu dem dynamischen Charakter beiträgt, da beispielsweise ein Surrogat zu einem bestimmten Zeitpunkt zu einer Suchanfrage relevant sein kann, und zu einem späteren Zeitpunkt weniger oder nicht relevant aufgrund der sich geänderten Situation und des Kontexts der informationssuchenden Person.

Zusammenfassend lässt sich Relevanz beschreiben als dynamisch, multidimensional und kontextabhängig (situationsabhängig), dabei ist zu beachten: „One has to realize that dynamic aspects of relevance are quite different than situational aspects – former relate to the process and later to the context“ (Saracevic, 2012, S. 57). Diese Unterscheidung zwischen dem Prozess der Bewertung und dem Kontext der Bewertung ist von besonderer Bedeutung und wird im nachfolgenden Abschnitt weiter aufgegriffen, indem die Relevanzbewertung als Prozess des Urteilens definiert wird.

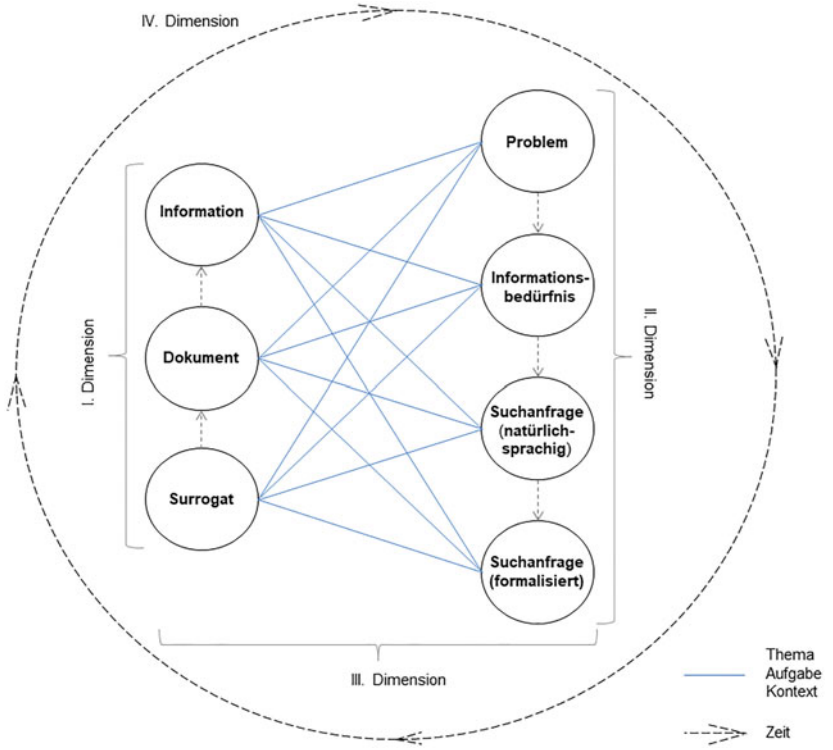

Abbildung 3.2 Multidimensionalität von Relevanzbeziehungen nach Mizzaro (1997) (modifizierte Darstellung)

3.1.3 Relevanzbewertung als Prozess des Urteilens

Wang (1994) kennzeichnet im Rahmen ihrer Studie die Relevanzbewertung als einen im Prozess der Dokumentenauswahl eingebetteten Prozess, wodurch dieser zugleich ein Teilprozess des übergeordneten menschlichen Informationssuchprozesses ist (vgl. Abschnitt 2.1.4). Wang (1994) betrachtet in ihrer Arbeit den Prozess der Dokumentenauswahl unter entscheidungstheoretischen Gesichtspunkten, was sie zwar nicht explizit begründet – „… *it may be worthwhile to look at an individual's document selection behavior from the point of view of decision making*" (Wang, 1994, S. 7) – in der gemeinsamen Publikation ihrer Studie

mit Soergel wird der Prozess der Dokumentenauswahl jedoch explizit als Entscheidungsprozess definiert, wobei die Auswahl nach einem *predictive judgment* erfolgt:

> Document selection is a decision process in which the user evaluates a retrieved document based on its surrogate obtained from a bibliographic IR system to decide whether or not to further pursue the document: to browse the actual document, to obtain a copy of the document, to read the document, or to make a reference to the document. (Wang & Soergel, 1998, S. 115)

Die Theorie des Entscheidens (*decision making*) bzw. das Entscheiden ist ein in der Emotions- und Motivationspsychologie bzw. kognitiven Psychologie angesiedeltes Forschungsfeld. Das Entscheiden ist dort wie folgt definiert:

> Prozess des Wählens zw. mind. zwei Optionen, mit dem Ziel […] erwünschte Konsequenzen zu erreichen und unerwünschte Konsequenzen zu vermeiden. Der Prozess führt im günstigen Fall zu einer Entscheidung. Durch die Entscheidung wird eine Option ausgewählt und der Entschluss (die *Intention*) gebildet, diese zu realisieren, z.B. indem eine *Handlung* ausgeführt wird. E. wird i.d.R. dem Forschungsfeld des *Judgment and Decision Making (JDM)* zugeordnet. Dort wird allerdings nicht immer klar zw. Urteilen und E. unterschieden. E. geht aber i.d.R. über *Urteilen* hinaus, da sich E. im Unterschied zu Urteilen auf die Bildung einer Handlungsintention bezieht und damit direkt handlungsbestimmend ist. Nichtsdestotrotz beruht E. häufig auf Urteilen, bes. über den Wert von Handlungsoptionen und die *Wahrscheinlichkeit*, dass diese eintreffen… (Plessner, 2017a; Kursivdruck im Original)

Bezüglich des Prozesses der Dokumentenauswahl betreffen die nach obiger Definition mindestens zwei Optionen die Entscheidung, ob ein Dokument ausgewählt wird oder nicht, wie bei Wang und Soergel gekennzeichnet – „*whether or not to further pursue the document*“ (Wang & Soergel, 1998, S. 115). Dieser Entscheidung muss die Bewertung des Dokuments zwangsläufig vorausgehen. Wenn der Prozess der Dokumentenauswahl einen Entscheidungsprozess darstellt, stellt sich die Frage, ob der Prozess der Relevanzbewertung ebenfalls als ein Entscheidungsprozess zu betrachten ist. Um diese Frage zu beantworten, ist die nähere Betrachtung des Konzepts des Urteilens hilfreich.

Urteilen ist ebenso im Bereich der Emotions- und Motivationspsychologie sowie in der kognitiven Psychologie verortet. Beim Urteilen handelt es sich um den psychologischen Prozess,

> …der zugrunde liegt, wenn Menschen einem Urteilsobjekt (z.B. einer Person, einem Gegenstand oder einer Aussage) einen Wert auf einer Urteilsdimension zuordnen

> (z.B. von sehr gut bis sehr schlecht oder wahr/falsch) und das daraus resultierende Urteil explizit zum Ausdruck bringen. […Urteilen] ist zum einen zu unterscheiden vom Prozess des Wahrnehmens […] und zum anderen vom Entscheiden […] U. kann wiederum die Grundlage für Entscheiden sein, allerdings führt U. nicht notwendigerweise wie Entscheiden zu Handlungen… (Plessner, 2017b).

Der Prozess der Relevanzbewertung von Surrogaten kann demnach als ein Prozess des Urteilens und nicht als Entscheidungsprozess verstanden werden; das Ergebnis des Bewertungsprozesses ist das Urteil über das Surrogat, auf dessen Basis die Entscheidung für oder gegen die Dokumentenauswahl getroffen wird (ob ein Suchergebnis angeklickt wird oder nicht).[5] Das Urteil zum Ausdruck zu bringen, entspricht im Forschungskontext bei der Erhebung von expliziten Relevanzbewertungen diese anhand einer Skala zu kennzeichnen oder beispielsweise unter Anwendung der Methode des lauten Denkens entsprechend zu verbalisieren. In der Literatur taucht mitunter der Begriff Relevanzentscheidung (*relevance decision*) auf, der nach obiger Darlegung nicht korrekt ist, weil er die Relevanzbewertung mit dem Produkt eines Entscheidungsprozesses gleichsetzt; dahingehend wäre der Begriff Relevanzurteil ein angemessenes Synonym zur Relevanzbewertung im Sinne eines Produktes im Prozess der Erhebung von expliziten Relevanzbewertungen.

Abbildung 3.3 veranschaulicht die Unterscheidung zwischen der Relevanzbewertung als Produkt des Urteilens und der Dokumentenauswahl als Ergebnis des Entscheidens und folgt somit dem Ansatz von Wang (1994), den Prozess der Relevanzbewertung als eingebettet in den Prozess der Dokumentenauswahl zu definieren.

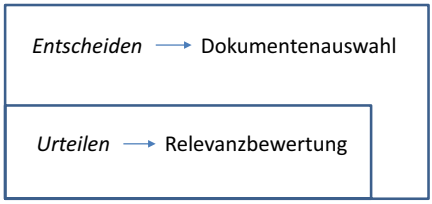

Abbildung 3.3
Relevanzbewertung und Dokumentenauswahl

[5] Vor diesem Hintergrund wird deutlich, dass Klickdaten aufgrund ihrer dichotomen Ausprägung eine Entscheidung und keine Beurteilung darstellen, wodurch sie als Alternative zur Erhebung expliziter Relevanzbewertungen nicht geeignet sind.

3.2 Identifikation der Einflüsse im Prozess der Relevanzbewertung von Surrogaten

Das Ziel dieses Abschnitts ist die Darstellung der unterschiedlichen Einflüsse im Prozess der Relevanzbewertung und damit einhergehend die Beantwortung der Forschungsfragen F1a (*Wie lassen sich Merkmale, Kriterien und Faktoren als Einflüsse im Prozess der Relevanzbewertung für die Entwicklung eines experimentellen Untersuchungsdesigns definitorisch und konzeptuell voneinander abgrenzen?*) und F1b (*Wie können Kriterien bei der Relevanzbewertung von Suchergebnissen für eine experimentelle Studie operationalisiert werden?*). Zu diesem Zweck wird ein Modell entwickelt, welches das Wirken dieser unterschiedlichen Einflüsse grafisch darstellt. Das Modell bietet zugleich einen Lösungsvorschlag des Definitionsproblems in Hinblick auf die uneindeutigen Grenzen der Begriffe Kriterien (*criteria*), Faktoren (*factors*) und Merkmale (*clues/cues*) und deren inkonsistente Verwendung im Kontext der Erforschung von Relevanzkriterien in der Literatur.

Das Modell greift inhaltlich die Ergebnisse der Literaturschau auf. Insbesondere das Modell zur Bewertung der Informationsqualität und Autorität von Rieh (vgl. Abschnitt 2.1.2.1) und das Modell zur Dokumentenauswahl von Wang (vgl. Abschnitt 2.1.4) inspirierten die Darstellung der Einflussparameter bei der Relevanzbewertung. Zusätzlich werden die Elemente in Surrogaten berücksichtigt, die heutzutage von modernen akademischen Suchsystemen in der Suchergebnisliste integriert sind und der informationssuchenden Person Hinweise auf die Relevanz des Suchergebnisses liefern. Konkret bedeutet dies, dass das Modell zusätzlich Popularitätsdaten als expliziten Bestandteil eines Surrogats mit einbezieht (vgl. Abschnitt 2.1.3). Solche zusätzlichen Daten waren in den bisherigen Studien, in denen Surrogate als Untersuchungsgegenstand dienten, nicht enthalten.

Hinzu kommt, dass in bisherigen Studien zu Relevanzkriterien nicht immer explizit zwischen *predictive judgments* und *evaluative judgments* unterschieden wurde (vgl. Abschnitt 2.1.4). Je nachdem, welchen Umfang das bewertete Informationsobjekt hat, d. h. ob es sich um ein Surrogat oder den eigentlichen Inhalt (Volltext) handelt, sind andere Elemente enthalten, auf Basis derer Relevanzbewertungen vorgenommen werden. Diese Unterscheidung wird in dem Modell ebenfalls berücksichtigt, da es auf die Darstellung der Einflüsse bei der Surrogatbewertung, also auf die Anwendung von *predictive judgments* durch die informationssuchende Person, abzielt.

Neben der zuvor vorgestellten ziel- und aufgabenorientierten Definition von Relevanz (vgl. Abschnitt 3.1.1), liegen dem Modell die folgenden Annahmen zugrunde:

- Das Modell zeigt den Prozess der Relevanzbewertung von Surrogaten, die als Ergebnis einer informationsorientierten im Gegensatz zu einer navigations- oder transaktionsorientierten Suchanfrage (vgl. Broder, 2002) vom IR-System ausgegeben werden, da bei letztgenannten die Relevanzbewertung mit anderen Kriterien einhergehen würde, weil beispielsweise Suchergebnisse zu navigationsorientierten Anfragen keine Relevanzbewertungen in diesem Sinne verlangen, sondern lediglich zu prüfen ist, ob das eine gesuchte Informationsobjekt tatsächlich gefunden wurde, wobei diese Prüfung anhand eines reinen Text matching erfolgen kann (vgl. Abschnitt 1.1). Konkret lautet die Annahme, dass es sich nicht um die Suche nach einem bereits bekannten Dokument handelt.
- Wenn für informationsorientierte Suchanfragen nach wissenschaftlichen Informationen IR-Systeme genutzt werden, handelt es sich in der Regel um akademische Suchsysteme.
- Akademische Suchsysteme werden von informationssuchenden Personen genutzt, die einen akademischen Hintergrund bzw. Kontext aufweisen, wie beispielsweise Studierende, Hochschullehrende, Promovierende, Forschende.
- Das Modell geht von einem textbasierten IR-System aus, das eine auf Basis eines Rankingalgorithmus sortierte Trefferliste auf eine Suchanfrage ausgibt.
- Die angezeigten Suchergebnisse enthalten die bibliografischen Angaben zu den Dokumenten unabhängig von einem direkten Link auf den Volltext.

Abbildung 3.4 zeigt das Modell zur Relevanzbewertung aus der Perspektive einer informationssuchenden Person, die mit dem akademischen Suchsystem interagiert, und veranschaulicht den Prozess der Bewertung von Surrogaten, der auch als *predictive judgment process* (Balatsoukas & Ruthven, 2010, S. 1) bzw. *process of predictive judgment* (Crystal & Greenberg, 2006, S. 1369) bezeichnet werden kann. Merkmale stellen die Elemente oder Attribute in einem Surrogat dar und dienen als Hinweisreize für die informationssuchende Person, von der sie wahrgenommen werden.[6] Ausgehend von diesen Merkmalen werden die Kriterien gebildet, die zusätzlich in irgendeiner Weise gewichtet werden. Diese Verarbeitung von Surrogatelementen und die Bildung und Gewichtung von Kriterien durch die informationssuchende Person *bilden* den Prozess der Relevanzbewertung; Faktoren hingegen *beeinflussen* diesen Prozess bezüglich des Kontexts der Situation, in der sich die informationssuchende Person zum Zeitpunkt der Interaktion

[6] Diese Anerkennung der Elemente eines Surrogats als Hinweisreize findet sich bereits bei Wang (1994), die auf das Linsenmodell von Egon Brunswick zurückgreift (vgl. Abschnitt 2.1.4, S. 40).

mit dem Suchsytem befindet. Ist von Einflussfaktoren bei der Relevanzbewertung die Rede, sollten demnach nicht die der Bewertung zugrunde liegenden subjektiven Kriterien gemeint sein, obwohl diese auch Einfluss auf das Urteil (Produkt des Bewertungsprozesses) nehmen, sondern die Faktoren, die auf diesen Bewertungsprozess einwirken. Dieser Prozess endet entweder mit dem Zugriff auf den Volltext, den eigentlichen Inhalt, der die Voraussetzung für ein *evaluative judgment* ist, oder der Such- bzw. Bewertungsprozess wird fortgesetzt, wenn das Urteil nicht zu der Entscheidung für das Aufrufen des Volltextes führt.

Im Modell ist die Gewichtung der Kriterien im Kontext der Urteilsbildung getrennt von der Kriterienbildung dargestellt, sie erfolgt unmittelbar vor der Entscheidung zur Dokumentenauswahl, welche entweder negativ (Dokument wird nicht ausgewählt, nicht geklickt) oder positiv (Dokument wird ausgewählt, geklickt) ausfällt. An dieser Stelle wird die Bedeutung der im vorangegangenen Abschnitt erläuterten Argumentation, den Prozess der Relevanzbewertung als einen Prozess des Urteilens zu betrachten, deutlich (vgl. Abschnitt 3.1.3): Mit der Unterscheidung zwischen Urteilen und Entscheiden wird die Relevanzbewertung begrenzt auf die Anwendung und Gewichtung von Relevanzkriterien, die schließlich zu einer Entscheidung führen, die Entscheidung erfolgt daher auf der Basis der Relevanzbewertung, die ihr zeitlich vorausgeht. In welcher Weise die Kriteriengewichtung erfolgt, die zur Dokumentenauswahl führt, bleibt abgesehen von der thematischen Relevanz als wichtigstes Kriterium unklar und wird in den nachfolgenden Abschnitten nicht weiterverfolgt.

Zusammengefasst sind die drei zentralen Aspekte, die das Modell in ihrem Zusammenspiel darstellt, (a) die Attribute des Surrogats, die als potenzielle Relevanzmerkmale betrachtet und als operationalisierte Kriterien definiert werden, (b) Kriterien für die Relevanzbewertung, die auf der Basis der Surrogatelemente gebildet werden, und (c) die Faktoren, die Einfluss auf den Bewertungsprozess nehmen und somit den Kontext der Situation, in der sich die informationssuchende Person zu dem jeweiligen Zeitpunkt befindet, abbilden. Diese werden in den beiden nachfolgenden Abschnitten vorgestellt. Zu beachten ist dabei, dass keine ausführliche Darstellung einzelner Surrogatelemente, Nutzerkriterien und Einflussfaktoren erfolgen kann, da für die wissenschaftliche Erörterung jedes einzelnen Elements, Kriteriums, Faktors eine separate und umfassende Literaturschau erforderlich wäre. Eine detaillierte Auseinandersetzung mit der entsprechenden Literatur würde den Rahmen dieser Arbeit sprengen und ist zudem für die Bearbeitung der weiteren Forschungsfragen nicht erforderlich.

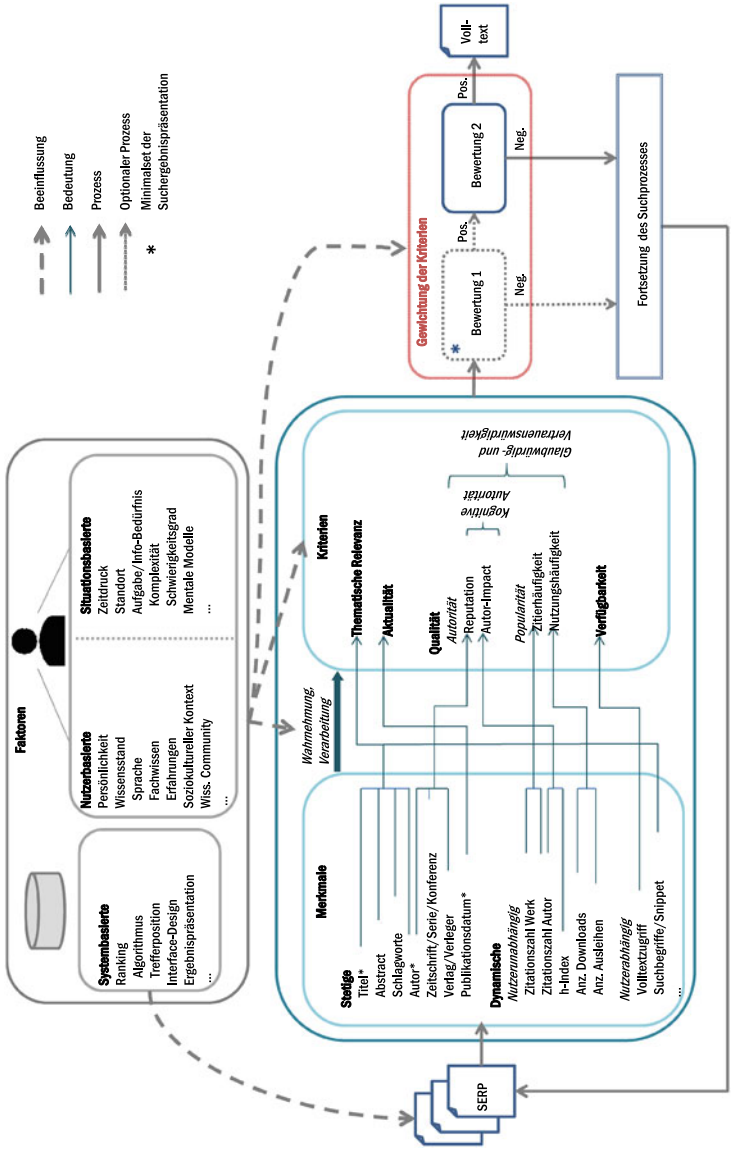

Abbildung 3.4 Modell zur subjektiven Relevanzbewertung von Suchergebnissen in akademischen Suchsystemen (basierend auf Behnert, 2019)

3.2.1 Attribute des Surrogats als Basis für die Kriterienbildung

Diverse Kriterien beeinflussen informationssuchende Personen im Prozess der Relevanzbewertung; thematische Relevanz gilt dabei als das wichtigste Kriterium, auf dessen Basis weitere Kriterien Anwendung finden (vgl. Abschnitt 2.1.1). Diese Relevanzkriterien lassen sich in vier Hauptkriterien gliedern: Neben der thematischen Relevanz sind diese Aktualität, Qualität und Verfügbarkeit (vgl. Abbildung 3.4). Bei der Qualität handelt es sich genau genommen um die erwartete oder vermutete Qualität, die wiederum auf der Basis von Autoritäts- und Popularitätskriterien abgeleitet wird. Autoritätskriterien wie Reputation und Autor-Impact können Hinweise auf potenzielle kognitive Autoritäten liefern, die gemeinsam mit Popularitätskriterien wie Zitier- und Nutzungshäufigkeit die Kriterien Glaubwürdigkeit und Vertrauenswürdigkeit bilden, welche insbesondere vor dem Hintergrund der Dynamik im Web und dessen exponentiellen Wachstums wesentliche Kriterien darstellen (vgl. Abschnitt 2.1.2).

Die Bildung und Anwendung der Relevanzkriterien erfolgt anhand der Attribute des Surrogates, die von der informationssuchenden Person wahrgenommen und verarbeitet werden. Diese Surrogatelemente dienen als Relevanzmerkmale und können als operationalisierte Kriterien betrachtet werden. Obwohl weitere Aspekte wie die Position des Treffers in der Suchergebnisliste ebenfalls wahrgenommen werden, zählen diese nicht zu den Merkmalen, sondern zu den Relevanzfaktoren, wie im nachfolgenden Abschnitt 3.2.2 erläutert wird. Diese Abgrenzung zwischen Merkmalen und Faktoren ist damit zu begründen, dass zwischen der Ebene des einzelnen Surrogats als Lieferant von Relevanzmerkmalen und der Ebene der Suchergebnisliste als von (systembasierten) Faktoren beeinflusste Ebene zu unterscheiden ist. Dadurch wird eine klare Trennung der Konzepte begünstigt.

Im Modell sind die Elemente eines Surrogates, die die potenziellen Relevanzmerkmale darstellen, in zwei Gruppen unterteilt: (a) Stetige Merkmale sind die formalen Metadaten eines Werkes, die unveränderlich und unabhängig von der Suchanfrage und dem Suchsystem sowie dem Kontext der informationssuchenden Person sind, wie Titel, Autor, Abstract; (b) Dynamische Merkmale sind Informationen, die sich im Lauf der Zeit verändern können. Bei diesen nicht fixen Merkmalen kann zudem zwischen nutzerunabhängigen und nutzerabhängigen Elementen unterschieden werden. Unabhängig von der informationssuchenden Person und ihrer Situation sind Angaben zu dem Dokument (Surrogat) wie Zitationszahlen und Informationen darüber, wie häufig ein Dokument bereits heruntergeladen oder – bezogen auf den Bibliothekskontext – ausgeliehen wurde.

Als nutzerabhängig sind die Elemente zu verstehen, die bezogen auf das jeweilige Suchergebnis und dessen Quelle verschiedene Angaben besitzen. So kann der physische oder virtuelle Standort der Person bzw. ihre Zugehörigkeit zu einer Einrichtung über den Zugang zu lizensierten Quellen entscheiden; des Weiteren werden die verwendeten Suchbegriffe, die mit Begriffen im Surrogat oder Snippet übereinstimmen, abhängig von der Suchanfrage entsprechend hervorgehoben.

Anhand welcher Elemente welche Kriterien abgeleitet werden können, ist in der Darstellung des Modells in Abbildung 3.4 mithilfe von Bedeutungspfeilen zwischen den Merkmalen und den entsprechenden Kriterien gekennzeichnet: Thematische Relevanz kann anhand von inhaltsbezogenen Merkmalen wie Titel und Abstract aber auch mithilfe von hervorgehobenen Suchbegriffen ermittelt werden; Aktualität ist hier als ein Kriterium gelistet, das ausschließlich auf dem Publikationsdatum beruht und unabhängig von der Aktualität des Inhalts beispielsweise im Sinne einer aktuell geführten gesellschaftlichen Debatte zu verstehen ist, die eher bei der thematischen Relevanz zu verorten wäre. Das Kriterium der Verfügbarkeit ist neben dem der Aktualität das am einfachsten abzuleitende Kriterium, für das im Prinzip keine kognitiven Ressourcen aufgebraucht werden müssen. Demgegenüber stehen die Kriterien für Qualität: Um die Reputation einer Autorin oder eines Autors, einer Zeitschrift oder Konferenz als Publikationsorgan, eines Verlages einschätzen zu können, sind Kenntnisse über diese, auch durch persönliche Erfahrungen, erforderlich. Ähnliches gilt für das Kriterium Autor-Impact, für das exemplarisch im Modell der h-Index als dynamisches, nutzerunabhängiges Merkmal gewählt wurde, das zwar auch Wissen um das Zustandekommen dieser Kennzahl benötigt, allerdings als explizite und quantifizierte Kennzahl der informationssuchenden Person keinen Interpretationsspielraum lässt – unabhängig von einer Bewertung dieser Kennzahl.

Die Popularitätskriterien Zitierhäufigkeit und Nutzungshäufigkeit basieren auf den bereits in Abschnitt 2.1.3 vorgestellten Popularitätsdaten, der Zitationszahl eines Werkes oder eines Autors/einer Autorin sowie der Anzahl der Downloads eines Werkes oder der Anzahl von Ausleihen im Kontext wissenschaftlicher Bibliotheken.

In der Realität sind nicht immer alle im Modell aufgeführten Merkmale in einem Surrogat enthalten. Die Darstellung kann nach System bzw. Anbieter oder der Art des Suchsystems (elektronisch oder analog, Bibliotheksbestand oder nicht) variieren. Manche Daten werden nach Kenntnisstand der Autorin aktuell nicht in die Suchergebnispräsentation integriert, z. B. der h-Index. Allerdings ist von einem Minimalset an präsentierten Elementen auszugehen, welches Titel, Autoren und Publikationsdatum beinhaltet. Ein solches Minimalset erfordert eine

erste Bewertung, um in einem zweiten Schritt weitere Elemente wie das Abstract zu erhalten (Volltrefferanzeige).

3.2.2 Einflussfaktoren als Kontext der Relevanzbewertung

In diesem Abschnitt wird die Rolle der den Relevanzbewertungsprozess beeinflussenden Faktoren dargestellt. Neben diesen Faktoren konnten in Studien weitere Einflussparameter aufgedeckt werden, die einen Einfluss auf die Bewertung haben und durch deren Kenntnis der Prozess der Relevanzbewertung besser verstanden werden kann. Bei diesen Effekten handelt es sich beispielsweise um Positions- bzw. Reihenfolgeneffekte, die insbesondere bei der Erhebung von expliziten Relevanzbewertungen im Forschungskontext berücksichtigt werden müssen. Solche unerwünschten Effekte und Verfahren zu ihrer Vorbeugung bei der Planung und Durchführung empirischer und insbesondere experimenteller Studien werden in Abschnitt 4.1.5.2 beleuchtet.

Die nachfolgende Vorstellung der Relevanzfaktoren erfolgt jedoch unabhängig von Einflüssen im Rahmen von Studien. Sie dient vorranging der Veranschaulichung der definitorischen Abgrenzung von Faktoren zu Kriterien anhand exemplarisch ausgewählter Faktoren, die auch generell als Einflussfaktoren auf das menschliche Suchverhalten und Suchstrategien im Informationssuchprozess bekannt sind.

In der Darstellung in Abbildung 3.4 wird die Unterteilung der Faktoren in systembasierte Faktoren und in Faktoren, die die informationssuchende Person und ihren Kontext betreffen, ersichtlich. Diese menschlichen Faktoren können wiederum in nutzerbasierte und situationsbasierte Faktoren gruppiert werden.

Als *systembasierte Faktoren* werden diejenigen Faktoren bezeichnet, die systembedingt wirken und Einfluss auf die Suchergebnispräsentation nehmen. Zunächst ist hier das Ranking zu nennen, das bereits in zahlreichen Studien als bedeutsamer Orientierungsfaktor nachgewiesen wurde. Die Treffersortierung wird durch den Ranking-Algorithmus bestimmt, der eine Vielzahl an Signalen auswertet, wie beispielsweise nutzerbasierte Daten für die Bestimmung von der Popularität und Glaubwürdigkeit von Dokumenten (vgl. Abschnitt 2.1.2).

Auch die Reihenfolge der Suchtreffer beeinflusst die Relevanzbewertung, welche abhängig von den Dokumenten ist, die bereits zuvor gesehen und beurteilt wurden. Dies wurde in Studien zum sogenannten *Order Effect* untersucht (vgl. z. B. Shokouhi, White, & Yilmaz, 2015; Xu & Wang, 2008), wie in Abschnitt 4.1.5.2 im Zusammenhang mit Effekten bei der Erhebung von

Relevanzbewertungen näher beschrieben wird. Die Art der Präsentation der Suchergebnisse stellt ebenso einen Relevanzfaktor dar. Traditionell werden Ergebnisse als eine gerankte Liste von dem IR-System ausgegeben, jedoch konnten Kammerer & Gerjets (2014) zeigen, dass die Bewertung der Vertrauenswürdigkeit von Quellen bei der Websuche positiv beeinflusst wird durch die Darstellung der Suchergebnisse als Raster (3×3) im Gegensatz zur üblichen Listendarstellung. Diese systembasierten Faktoren sind nicht beschränkt auf die Relevanzbewertung von Suchergebnissen in akademischen Suchsystemen, sondern können auch als Faktoren in nicht akademischen Suchsystemen, wie populäre Websuchmaschinen, Berücksichtigung finden.

Zu den *nutzerbasierten Faktoren* zählen solche Faktoren, die der informationssuchenden Person innewohnen und teilweise mehr oder weniger unveränderbar sind (z. B. das Erlernen einer Sprache). Von diesen grenzen sich situationsbasierte Faktoren ab, die zwar ebenfalls auf die informationssuchende Person bezogen sind, allerdings auch kurzfristig verändert werden können (z. B. den Wechsel des Standorts).

Die Literatur zeigt, dass die Persönlichkeit eines Menschen sein Verhalten bei der Informationssuche beeinflussen kann (Heinström, 2003, 2005); auch die Anwendung von Kriterien bei der Relevanzbewertung kann unter Umständen durch bestimmte Persönlichkeitsmerkmale beeinflusst werden (Sims, 2002). Dass das (Vor-)Wissen einer informationssuchenden Person maßgeblich ist für die Anwendung des Relevanzkriteriums der thematischen Relevanz, wurde bereits hervorgehoben. Auch wurde beschrieben, dass das Ableiten von thematischer Relevanz kognitive Ressourcen erfordert und über die Bestimmung der *Aboutness* von Dokumenten hinausgeht (vgl. Abschnitt 1.1 und Abschnitt 2.1.1). Der Wissensstand einer Person kann daher als bedeutendster Faktor bei der Relevanzbewertung gesehen werden. Darunter fallen Sprachkenntnisse, Wissen über ein bestimmtes Thema sowie spezielles Fachwissen (Fachbegriffe) und persönliche Erfahrungen. Bereits in den bekannten früheren Studien zu Relevanzkriterien wurde die Bedeutung des Wissensstands einer Person im Prozess der Relevanzbewertung deutlich. So schlussfolgert Barry (1998): „It seems clear that respondents were using their knowledge about previous work from sources to make predictions about the content and quality of any other work coming from those sources“ (S. 1302). Vakkari & Hakala (2000) fanden heraus, dass domänenspezifisches Wissen den Teilnehmenden ihrer Studie bei der Beurteilung von neu hinzu gekommenen Informationen half.

Außerdem sollte der soziokulturelle Kontext der informationssuchenden Person als Einflussfaktor verstanden werden, da dieser beispielsweise die Einstellung und das Verhalten gegenüber Autoritäten prägt, wie die Forschung von Hofstede

et al. (2017) zum Konzept der Machtdistanz (*Power Distance*) im Bereich interkultureller Zusammenarbeit aufzeigt: Demnach bestehen Unterschiede diesbezüglich beispielsweise bei manchen westlichen Ländern wie Schweden (geringe Machtdistanz) im Vergleich zu asiatischen Ländern wie China (große Machtdistanz). Bezüglich des Relevanzkriteriums Autor-Impact und dem Konzept der kognitiven Autorität wären in diesem Bereich weitere Untersuchungen von Interesse.

Einen weiteren Faktor stellt die eigene wissenschaftliche Community dar, deren Vorgaben zum Zitations- und Publikationsverhalten die Bewertung von Suchergebnissen im akademischen Kontext beeinflussen können. Wenn in den Geisteswissenschaften die Zitationszahl eine eher untergeordnete Rolle einnimmt im Vergleich beispielsweise zur Informatik, erscheint es erforschungswürdig, ob entsprechende Erwartungshaltungen beispielsweise hinsichtlich der Zitationszahl eines Werkes auf die Relevanzbewertung von Surrogaten, die solche Informationen beinhalten, übertragen werden.

Situationsbasierte Faktoren beziehen sich auf den Kontext der informationssuchenden Person und sind abhängig von der Situation, in der sie sich zum Zeitpunkt des Bewertungsprozesses befindet. Hier sind die Faktoren Zeitdruck oder Zeitvorgabe für die Bearbeitung einer Aufgabe oder die Lösung eines Informationsproblems zu nennen. Der Standort einer Person kann das Kriterium Verfügbarkeit beeinflussen, welches unter Umständen überhaupt keine Rolle spielt, wenn der Person beispielsweise über das eigene Hochschulnetz der sofortige Zugriff auf lizenzierte Dokumente ermöglicht wird. Da die Relevanzbewertung stets in Bezug auf die Befriedigung eines Informationsbedürfnisses erfolgt, sind der Schwierigkeitsgrad einer Suchaufgabe und ihrer Komplexität gewichtige Faktoren. Hier sind exemplarisch die Arbeiten von Byström & Hansen (2005), Hjørland & Christensen (2002) und Xie (2008) zu nennen, die im Kontext der Interactive Information Retrieval-Forschung ein breites Forschungsinteresse bedienen.

Schließlich sind als situationsabhängige Faktoren die mentalen Modelle und Heuristiken der Person zu nennen, da diese gewisse Erwartungen über die Ergebnisse, die als hochrelevant gelten können, hervorrufen. Aber auch Erwartungen (und die Erfahrungen) mit dem Suchsystem sind an dieser Stelle zu nennen: In ihren Experimenten untersuchte Werner (2019) unter anderem die Benutzererwartung auf die Relevanzwahrnehmung bei der Informationssuche und konnte nachweisen, dass eine positive Erwartungshaltung mit weniger restriktiv angewendeten Relevanzkriterien einhergeht. Dies ist ein Beispiel für den Einfluss der Faktoren auf die Gewichtung der Kriterien.

3.3 Zusammenfassung und Beantwortung der Forschungsfragen F1a & F1b

In diesem Kapitel wurden die Voraussetzungen zur experimentellen Erforschung von Relevanzkriterien geschaffen, indem zunächst das Relevanzkonzept spezifiziert und in diesem Zusammenhang die für die vorliegende Arbeit gewählte ziel- und aufgabenorientierte Relevanzdefinition erläutert wurde. Der Prozess der Relevanzbewertung von Suchergebnissen wurde als ein Prozess des Urteilens beschrieben. Diese Betrachtung ist in Hinblick auf die Definition des Begriffs Relevanzkriterium von besonderer Bedeutung, da die Kriterien für die Relevanzbewertung im Sinne einer Urteilsbildung abgegrenzt sind von der Gewichtung, die schließlich zu einer Entscheidung für oder gegen die Auswahl des Dokuments führt. Das Verständnis dieser Unterscheidung hilft dabei, die verschiedenen Einflüsse im Prozess der Relevanzbewertung zu identifizieren und definitorisch voneinander abzugrenzen. Die Identifikation der Einflüsse erfolgte auf Basis der Erkenntnisse der Literaturschau und führte zu der Entwicklung eines Modells zur subjektiven Relevanzbewertung von Suchergebnissen in akademischen Suchsystemen, welches die in heutigen Systemen integrierten Popularitätsdaten berücksichtigt. Anhand des Modells lässt sich die **Forschungsfrage F1a** (*Wie lassen sich Merkmale, Kriterien und Faktoren als Einflüsse im Prozess der Relevanzbewertung für die Entwicklung eines experimentellen Untersuchungsdesigns definitorisch und konzeptuell voneinander abgrenzen?*) beantworten: Auf Basis der in den Surrogaten enthaltenen Elemente als potenzielle Relevanzmerkmale leitet eine informationssuchende Person Relevanzkriterien ab, die auf eine bestimmte Weise gewichtet werden und zur Relevanzbewertung als Produkt führen (*predictive judgment*). Diese Bewertung ist die Voraussetzung für die Entscheidung, ein Dokument auszuwählen, d. h. dessen Volltext aufzurufen, oder es nicht auszuwählen. Das Aufrufen des Volltexts würde zu einem *evaluative judgment* führen, dies ist jedoch nicht Gegenstand des Modells.

Das Zusammenspiel der beiden Aspekte Merkmale und Kriterien wird durch verschiedene Faktoren beeinflusst, die als systembasiert, nutzerbasiert und situationsbasiert unterschieden werden können. Wichtig hierbei ist die Trennung der Ebene des einzelnen Surrogats von der Ebene der Suchergebnisliste. Zum Beispiel lässt sich anhand des Abstracts (Merkmal) die thematische Relevanz (Kriterium) ableiten, wobei Fachwissen oder Sprachkenntnisse (Faktoren) diesen Prozess beeinflussen können. Systembasierte Faktoren wie der Ranking-Algorithmus beeinflussen lediglich das Zustandekommen der Ergebnisliste, jedoch nicht die Art oder Anzahl von Elementen des einzelnen Surrogats.

Die zentrale Erkenntnis aus dieser systematischen Übersicht und zugleich die Antwort auf die **Forschungsfrage F1b** (*Wie können Kriterien bei der Relevanzbewertung von Suchergebnissen für eine experimentelle Studie operationalisiert werden?*) ist, dass die Elemente des Surrogats als Relevanzmerkmale als operationalisierte Kriterien zu betrachten sind und somit diejenigen messbaren Variablen darstellen, die für eine empirische Untersuchung von Kriterien anhand eines experimentellen Designs unabdingbar sind. Das Modell kann demzufolge auch als „Hilfsmittel" für die Operationalisierung in künftigen Studien zur Erforschung von Relevanzkriterien herangezogen und weiterentwickelt werden. Ferner bietet es einen Lösungsvorschlag für das Definitionsproblem und erstmals eine systematische Übersicht über die Merkmale, Kriterien und Faktoren, die im Prozess der Relevanzbewertung von Surrogaten eine Rolle spielen. Insbesondere hilft das Modell zu verstehen, wie die Elemente eines Surrogats als potenzielle Merkmale für Relevanz mit Relevanzkriterien zusammenhängen.

In Anbetracht der Identifikation der Einflüsse im Bewertungsprozess von Suchergebnissen verliert das Phänomen Relevanz im Kontext der Informationssuche nicht an Komplexität. Mithilfe des Modells wurde jedoch erstmals ein klareres Bild des subjektiven Bewertungsprozesses gezeichnet.

4 Studie zur Untersuchung des Einflusses von Popularitätsdaten auf die Relevanzbewertung von Suchergebnissen in akademischen Suchsystemen

Die Literaturschau zum aktuellen Stand der Forschung hinsichtlich Relevanz und Relevanzkriterien zeigte, dass bisher keine Studien zur Erforschung von Relevanzkriterien veröffentlicht wurden, in denen Teilnehmende Suchergebnisse (Surrogate) mit integrierten Popularitätsdaten bewerteten. Neue Erkenntnisse der Relevanzforschung und im weiteren Sinne der Informationsverhaltensforschung sollten auf Untersuchungen beruhen, in denen Eigenschaften gegenwärtig existierender Suchsysteme mitberücksichtigt werden. Moderne akademische Suchsysteme wie Google Scholar oder die ACM Digital Library integrieren beispielsweise die Anzahl von Downloads oder die Anzahl der Zitationen eines Werkes.

Des Weiteren wurden in den bisherigen Studien hauptsächlich explorative Designs verwendet, um Kenntnisse über Relevanzkriterien zu erhalten. Um aussagekräftige Erkenntnisse über direkte Zusammenhänge zwischen den Elementen von Suchergebnissen als Träger von Relevanzmerkmalen und den Kriterien, anhand derer Nutzerinnen und Nutzer die Relevanz von Suchergebnissen beurteilen, zu erlangen, sind experimentelle Untersuchungen, wie sie in empirischen Studien anderer (sozialwissenschaftlicher) Fachdisziplinen zur Erforschung von Verhalten und Einstellungen von Menschen durchgeführt werden, besser geeignet (vgl. Abschnitt 2.2.4).

Der Zweck der in diesem Kapitel vorgestellten Studie ist es, den Einfluss von Popularitätsdaten als explizitem Bestandteil von Suchergebnissen in akademischen Suchsystemen auf die Relevanzbewertung experimentell zu untersuchen.

Ergänzende Information Die elektronische Version dieses Kapitels enthält Zusatzmaterial, auf das über folgenden Link zugegriffen werden kann https://doi.org/10.1007/978-3-658-37512-6_4.

C. Behnert, *Popularität und Relevanz in der Suche*,
https://doi.org/10.1007/978-3-658-37512-6_4

Hinsichtlich ihrer Methode beantwortet die Studie die Forschungsfrage F1 und bezüglich des Untersuchungsgegenstands – Suchergebnissen mit integrierten Popularitätsdaten im akademischen Kontext – dienen die Ergebnisse der Beantwortung der Forschungsfragen F2 und F3 (vgl. Abschnitt 2.3): Mithilfe eines Online-Fragebogens bewerteten Versuchspersonen Surrogate mit manipulierten Popularitätsdaten (Zitations- und Downloadzahlen) in Bezug auf ein Informationsbedürfnis, wobei Aufbau und Ablauf der Befragung einem experimentellen Versuchsplan folgten.

Ein Vorteil des Online-Experiments im Vergleich zu einem Laborexperiment liegt in der Realisierung eines wesentlich größeren Stichprobenumfangs, als in einem Laborexperiment in der Regel erzielt werden kann. Dies ist im Rahmen der hier vorgestellten Studie erreicht worden. So konnten für die vollständige Bearbeitung des Online-Fragebogens mehr als 700 Teilnehmende gewonnen werden, wobei die Daten von 627 Personen in die statistische Analyse eingingen. Diesbezüglich hebt sich diese Studie von den früheren Studien zu Relevanzkriterien deutlich ab.

Dieses Kapitel beginnt mit der Erläuterung der Methoden hinsichtlich des experimentellen Untersuchungsdesigns (Abschnitt 4.1), gefolgt von der Datenerhebung (Abschnitt 4.2) und der Datenanalyse (Abschnitt 4.3). Im Anschluss werden die Ergebnisse präsentiert (Abschnitt 4.4) und diskutiert (Abschnitt 4.5) sowie die Grenzen der Studie erläutert (Abschnitt 4.6).

Die in Abschnitt 2.3 gestellte Forschungsfrage F1 gibt bereits die Art des Untersuchungsdesigns vor, indem sie nach der Realisierbarkeit von Experimenten zur Erforschung von Relevanzkriterien fragt. Von der Art des Designs handelt es sich demnach bei der hier vorgestellten Studie um ein Experiment, bezüglich der Art der Datenerhebung um eine Befragung. Da die Befragung mithilfe eines Online-Fragebogens erfolgte, wird diese Studie auch als Online-Experiment bezeichnet. Für ein besseres Verständnis zur methodischen Verortung der Studie gibt Tabelle 4.1 einen schematischen Überblick über die Klassifikationskriterien sozialwissenschaftlicher Studiendesigns von Döring & Bortz (2016, S. 183) und ordnet die vorliegende Studie entsprechend ein. Konkret stellt sie demnach eine originale, experimentelle Stichprobenstudie ohne Messwiederholungen dar, deren Primärdaten im Feld[1] erhoben wurden.

Aufbauend auf den Erläuterungen zu experimentellen Studiendesigns in Abschnitt 2.2.4 wird in Abschnitt 4.1 die Entwicklung des Untersuchungsdesigns

[1] Empirische Studien, die ihre Daten nicht im Labor unter kontrollierten Laborbedingungen erheben, werden als Feldstudien bezeichnet; ein Online-Experiment, für das Versuchspersonen nicht in ein Labor geladen werden, gilt daher als Feldexperiment (Döring & Bortz, 2016, S. 207).

detailliert erläutert. Dabei wird zunächst auf die Auswahl und Operationalisierung der manipulierten unabhängigen Variablen (Abschnitt 4.1.1), der gemessenen abhängigen Variable (Abschnitt 4.1.2) und die Hypothesenentwicklung (Abschnitt 4.1.3) eingegangen. Anschließend werden weitere, möglicherweise als Ursache für die gemessenen Werte infrage kommenden Variablen, und der Umgang mit diesen potenziellen Störvariablen beschrieben (Abschnitt 4.1.4). Der Ablauf der Untersuchung erfolgte anhand eines Within-Subjects-Designs, in dem alle Versuchspersonen allen experimentellen Bedingungen ausgesetzt waren. Jede Person bearbeitete in jeweils randomisierter Reihenfolge dieselben Aufgaben vollständig (Abschnitt 4.1.5).

Abschnitt 4.2 beschreibt, wie die Daten mithilfe eines Online-Fragebogens erhoben wurden. Die Versuchspersonen bearbeiteten jeweils drei Aufgaben. Jede Aufgabe bestand in der Bewertung von neun Suchergebnissen zu einem vorgegebenen Informationsbedürfnis. Insgesamt wurden für das Experiment drei Aufgaben zur Relevanzbewertung entwickelt. Die Entwicklung der drei Informationsbedürfnisse (Abschnitt 4.2.1) sowie die Auswahl und Erstellung der insgesamt 27 Surrogate (Abschnitt 4.2.2) basierte auf einer Vielzahl von Entscheidungen, auf deren Begründungen ausführlich eingegangen wird. Neben den drei Bewertungsaufgaben erfasste der Fragebogen zusätzlich demografische Angaben der Teilnehmenden und Angaben zu Kenntnissen und der Nutzung wissenschaftlicher Suchsysteme; ebenso wurden eine Selbsteinschätzung über verwendete Bewertungskriterien und ein Meinungsbild über den Einfluss der untersuchten Popularitätsdaten im Anschluss an die Bewertungen erfasst. Aufgrund des Ziels dieser Arbeit, ein nachnutzbares methodisches Framework zu liefern, wird auch der Erläuterung der einzelnen Schritte der Fragebogenentwicklung besondere Aufmerksamkeit zuteil (Abschnitt 4.2.3). Die für eine valide Ergebnisse erzielende Erhebung erforderliche Anzahl an Versuchspersonen von $n = 577$ wurde vor Studienbeginn mit dem Statistik-Tool G*Power berechnet. Die dafür notwendigen Parameter und Einstellungen werden in Abschnitt 4.2.4 vorgestellt. Schließlich wird in Abschnitt 4.2.5 das Vorgehen bei der Gewinnung von Versuchspersonen erläutert.

Der darauffolgende Abschnitt 4.3 beschreibt die Datenanalyse von der Aufbereitung der erhobenen Rohdaten über die Auswahl des statistischen Verfahrens bis zur Erstellung des statistischen Modells mit der Software SPSS (Version 25).

Tabelle 4.1 Einordnung der Studie anhand der neun Klassifikationskriterien für Untersuchungsdesigns nach Döring & Bortz (2016, S. 183)

Kennzeichen des Untersuchungsdesigns	**Varianten von Untersuchungsdesigns**	**Diese Studie ist eine …**
1. Wissenschaftstheoretischer Ansatz der Studie	– Quantitative Studie – Qualitative Studie – Mixed-Methods-Studie	Quantitative Studie
2. Erkenntnisziel der Studie	– Grundlagenwissenschaftliche Studie – Anwendungsorientierte Studie a) Unabhängige Studie b) Auftragsstudie	Unabhängige, anwendungsorientierte Studie
3. Gegenstand der Studie	– Empirische Studie a) Originalstudie b) Replikationsstudie – Methodenstudie – Theoriestudie a) Review/Forschungsüberblick b) Metaanalyse	Empirische Originalstudie
4. Datengrundlage bei empirischen Studien	– Primäranalyse – Sekundäranalyse – Metaanalyse	Primäranalyse
Kennzeichen des Untersuchungsdesigns	Varianten von Untersuchungsdesigns	Diese Studie ist eine …

(Fortsetzung)

Tabelle 4.1 (Fortsetzung)

Kennzeichen des Untersuchungsdesigns	Varianten von Untersuchungsdesigns	Diese Studie ist eine …
5. Erkenntnisinteresse bei empirischen Studien	– Explorative (gegenstandsbeschreibende/theoriebildende) Studie – Deskriptive (populationsbeschreibende) Studie – Explanative (hypothesenprüfende) Studie	Explanative Studie
6. Bildung und Behandlung von Untersuchungsgruppen bei explanativen Studien	– Experimentelle Studie bzw. randomisierte kontrollierte Studie – Quasi-experimentelle Studie bzw. nicht randomisierte kontrollierte Studie – Nicht-experimentelle Studie	Experimentelle Studie
7. Untersuchungsort bei empirischen Studien	– Laborstudie – Feldstudie	Feldstudie
8. Anzahl der Untersuchungszeitpunkte bei empirischen Studien	– (Quasi-)Experimentelle Studie mit und ohne Messwiederholungen a) (Quasi-)Experimentelle Studie ohne Messwiederholungen b) (Quasi-)Experimentelle Studie mit Messwiederholungen – Nicht-experimentelle Studien mit und ohne Messwiederholungen a) Querschnittstudie b) Trendstudie c) Längsschnittstudie	Experimentelle Studie ohne Messwiederholungen
9. Anzahl der Untersuchungsobjekte bei empirischen Studien	– Gruppenstudie a) Stichprobenstudie b) Vollerhebung – Einzelfallstudie	Stichprobenstudie

Ein notwendiger Schritt in der Entwicklung einer empirischen Studie ist die Durchführung eines Pretests, mit dem der gesamte Versuchsablauf vorab mit einer kleineren Stichprobe getestet wird, um methodische Entscheidungen zu evaluieren und bei Bedarf zu optimieren (Döring & Bortz, 2016, S. 22 ff.). Dem für die hier beschriebene Studie durchgeführten Pretest wird für einen besseren Lesefluss kein eigener Abschnitt gewidmet, stattdessen fließen die Erkenntnisse aus dem Pretest in die jeweiligen Abschnitte ein.

Vorrangiges Ziel des Pretests war es, die Bearbeitungsdauer zu messen und die Verständlichkeit der Aufgabenstellungen zu überprüfen. Die Daten für den Pretest wurden vom 22. bis 31. März 2019 erhoben; Einladungen zur Teilnahme wurden per E-Mail an alle Wissenschaftlichen Mitarbeiterinnen und Mitarbeiter der HAW Hamburg versendet mit dem Hinweis, dass es sich um einen Pretest handelt und Feedback mithilfe der Kommentarfunktion des Online-Fragebogens ausdrücklich erwünscht ist. Die Anbindung der Autorin an die HAW Hamburg begründete die Erwartung, dass andere Promovierende sich solidarisch zeigen und darüber hinaus die Vergabe eines Amazon-Gutscheins im Wert von 20 EUR für jeden abschließend bearbeiteten Fragebogen einen finanziellen Anreiz zur Teilnahme bieten würde.

Die E-Mail-Adressen wurden über das hochschulinterne Mitarbeiterverzeichnis für jede Fakultät ermittelt[2]. Insgesamt wurden auf diese Weise 156 potenzielle Versuchspersonen erreicht, von denen 24 teilnahmen und schließlich 10 den Fragebogen tatsächlich beendeten, während 14 Teilnehmende vorzeitig abbrachen. Die Rücklaufquote der potenziellen Teilnehmenden lag somit bei 15,38 %, die Beendigungsquote bei 6,41 %. Die ausgewerteten Rückmeldungen der Teilnehmenden beschränkten sich dabei nicht auf diejenigen, die den Fragebogen bis zum Ende bearbeiteten; insbesondere lag das Interesse bei dem Feedback derjenigen, die die Bearbeitung abbrachen und zusätzlich in den Informationen über Abbruchseite und Bearbeitungsdauer. Die Erkenntnisse aus dem Pretest führten zu einer Anpassung des experimentellen Designs (vgl. Abschnitt 4.1.1) und damit einhergehend der Erstellung der Informationsbedürfnisse und Surrogate (Abschnitt 4.2.1 und Abschnitt 4.2.2) sowie des Fragebogenaufbaus (Abschnitt 4.2.3).

[2] Nicht alle an der HAW Hamburg Promovierenden sind zugleich als Wissenschaftliches Personal angestellt, daher wurde eine separate Anfrage an das Promotionszentrum mit der Bitte um Weiterleitung an alle über den E-Mail-Verteiler erreichbaren Promovierenden gestellt; eine Rückmeldung blieb allerdings aus. Weil dennoch genügend Personen an dem Pretest teilnahmen, wurde auf ein erneutes Anschreiben verzichtet.

4.1 Entwicklung des experimentellen Untersuchungsdesigns

Die Grundidee dieser Studie besteht darin, Versuchspersonen eine Reihe von Suchergebnissen vorzulegen und diese in Bezug auf ein vorgegebenes Informationsbedürfnis einzeln bewerten zu lassen. Die Suchergebnisse stellen das manipulierte Stimulusmaterial dar, welches die als unabhängige Variablen (UVn) manipulierten Popularitätsdaten enthält, um deren Effekt auf die Relevanzbewertungen, also die abhängige Variable (AV), zu messen. Methodisch folgt die Studie dem klassischen Ansatz bei Retrieval-Evaluierungen, bei dem den Studienteilnehmenden Suchaufgaben (Tasks) mit Kontextbeschreibungen (Beschreibungen von Informationsbedürfnissen) und Suchergebnissen in einer manipulierten Reihenfolge zur expliziten Relevanzbewertung vorgelegt werden, um anschließend relevanzbasierte Evaluierungskennzahlen zu berechnen und beispielsweise die verschiedenen Rankingalgorithmen miteinander zu vergleichen.

Zur Veranschaulichung der Wirkungsweise experimenteller Designs sollen derartige Retrieval-Studien als Beispiel dienen: Es wird vermutet, dass das Ranking diejenige unabhängige Variable (UV) ist, die eine beobachtbare Wirkung auf die Relevanzbewertung, also die abhängige Variable (AV), erzielt. Die Wirkung läge in einer höheren oder niedrigeren Relevanzbewertung einer festgelegten Anzahl von Dokumenten auf ausgewählten Positionen bzw. in einer höheren oder niedrigeren Kennzahl zur Feststellung der Retrieval-Effektivität. Die UV hat zwei verschiedene Ausprägungen, auch Stufen, die manipuliert sind: Ranking A und Ranking B. Diese stellen die zwei Bedingungen dar, denen die Probanden ausgesetzt sind. Dabei handelt es sich entweder um zwei Experimentalbedingungen oder um eine Kontrollbedingung (Baseline) und eine Experimentalbedingung. Nun können entweder die Probanden in zwei Gruppen aufgeteilt werden und nur jeweils die Aufgaben in einer der beiden Gruppen bearbeiten, also nur einer der beiden Bedingungen ausgesetzt sein; oder alle Probanden bearbeiten alle Aufgaben beider Gruppen und sind somit beiden Bedingungen ausgesetzt (Sedlmeier & Renkewitz, 2018, S. 154). In der erstgenannten Variante liegt ein Between-Subjects-Design vor: Die Stufen der UV werden zwischen den Probandengruppen variiert. In der letztgenannten Variante erfolgt die Erhebung anhand eines Within-Subjects-Designs: Jede/r Proband/in durchläuft beide Bedingungen bzw. Stufen der UV und bearbeitet alle Aufgaben beider Gruppen, d. h. die UV wird innerhalb der Probanden variiert (Sedlmeier & Renkewitz, 2018, S. 139).

Für das hier vorgestellte Online-Experiment wurde ein Within-Subjects-Design entwickelt. Ein solches Design ist bereits allein aufgrund des vorliegenden Untersuchungsgegenstands sinnvoll, da „subjektive Urteile über Merkmale von

Stimuli“ (Sedlmeier & Renkewitz, 2018, S. 157) untersucht werden, und „[s]olche Urteile über einen Stimulus hängen häufig von dem Kontext ab, in dem er dargeboten wird“ (Sedlmeier & Renkewitz, 2018, S. 157). Kontextabhängigkeit ist eine zentrale Komponente im informationswissenschaftlichen Relevanzkonzept (vgl. Abschnitt 3.1), zum Beispiel in Hinblick auf die Situation, in der sich die informationssuchende Person zum Zeitpunkt der Interaktion mit dem Suchsystem befindet, den Wissensstand der Person oder die konkrete Suchergebnisliste als Kontext eines einzelnen Suchergebnisses[3]. Demnach ist die Kontextabhängigkeit auch bei dem Prozess der Relevanzbewertung zu berücksichtigen; zudem ist der Prozess der Relevanzbewertung ein subjektives Urteilen (vgl. Abschnitt 3.1.3). Allein vor diesem Hintergrund stellt ein Within-Subjects-Design für diese Studie die richtige Wahl dar.

Hinzu kommen jedoch generelle Vorteile eines Within-Subjects-Designs gegenüber einem Between-Subjects-Design. So lassen sich auch kleine Effekte einer UV aufdecken (Sedlmeier & Renkewitz, 2018, S. 159), weil je Versuchsperson die Differenzen zwischen den einzelnen Bedingungen berechnet und dadurch Unterschiede pro Person von der Fehlervarianz getrennt werden können. Within-Subjects-Designs sind demnach sensitiver als Between-Subjects-Designs. In letzteren können aufgrund der Zuweisung der Teilnehmenden zu ausschließlich einer Bedingung keine Differenzen innerhalb einer Versuchsperson ermittelt werden (Sedlmeier & Renkewitz, 2018, S. 154).

Ein weiterer Vorteil gegenüber Between-Subjects-Designs liegt darin, dass alle personengebundenen Störvariablen (z. B. Alter, Einkommen, Motivation) parallelisiert sind, da jede Versuchsperson alle Bedingungen durchläuft. Somit ist eine mögliche Konfundierung dieser Störvariablen ausgeschlossen und weitere Kontrollmaßnahmen (vgl. Abschnitt 4.1.4) sind diesbezüglich nicht erforderlich (Sedlmeier & Renkewitz, 2018, S. 157).

Schließlich sind Within-Subjects-Designs auch aus forschungsökonomischen Gründen empfehlenswert, da sie eine geringere Anzahl an Versuchspersonen erfordern (Sedlmeier & Renkewitz, 2018, S. 157), wobei diese einen größeren Zeitaufwand aufbringen müssen als es mit einem Between-Subjects-Design der Fall wäre. Im oben genannten Beispiel gibt es zwei Gruppen, auf die in einem Between-Subjects-Design die Versuchspersonen verteilt werden, d. h. es werden zwei Teilstichproben miteinander verglichen. Wenn bereits eine weitere UV mit zwei Ausprägungen in das Design aufgenommen wird, handelt es sich

[3] Auf die Abhängigkeit der Bewertung eines Suchergebnisses von anderen Suchergebnissen einer Suchergebnisliste wird in Abschnitt 4.1.5.2 näher eingegangen.

um ein 2×2-Design[4] mit 4 experimentellen Bedingungen, sodass insgesamt vier Teilstichproben miteinander verglichen werden. Angenommen, pro Bedingung sollen die Daten von 50 Versuchspersonen erhoben werden, verlangt in diesem Fall ein Between-Subjects-Design insgesamt 200 Versuchspersonen, ein Within-Subjects-Design hingegen 50 Personen. In der hier vorgestellten Studie wäre ein Between-Subjects-Design in Hinblick auf die optimale Stichprobengröße sehr schwierig umzusetzen, da mit 27 experimentellen Bedingungen aus drei unabhängigen Variablen in jeweils drei Ausprägungen eine relativ große Zahl an Teilstichproben existiert.

Die Anzahl und Merkmale der für das hier beschriebene Design manipulierten und gemessenen Variablen, deren Skalenniveaus, die Hypothesenbildung sowie der Umgang mit Störvariablen (Drittvariablen) und die Planung des Versuchsaufbaus werden in den nachfolgenden Abschnitten erläutert.

4.1.1 Unabhängige Variablen

Wie im Rahmen der Identifikation der Einflüsse im Prozess der Relevanzbewertung dargelegt, verwenden informationssuchende Personen mehrere verschiedene Kriterien bei der Relevanzbewertung von Suchergebnissen (vgl. Abschnitt 3.2). Konkret ist davon auszugehen, dass auch verschiedene Arten von Popularitätsdaten als operationalisierte Popularitätskriterien einen ursächlichen Effekt auf die Relevanzbewertung von Surrogaten in wissenschaftlichen Suchsystemen erzielen (vgl. Abschnitt 4.1.3). Demzufolge sollten mindestens zwei Faktoren als mögliche Einflüsse untersucht werden. Für dieses Experiment wurden drei UVn ausgewählt. Das experimentelle Design ist somit ein dreifaktorielles Within-Subjects-Design.

In dem Modell zur subjektiven Relevanzbewertung von Suchergebnissen in akademischen Suchsystemen (vgl. Abschnitt 3.2) stellen Zitier- und Nutzungshäufigkeiten die Popularitätskriterien dar, die exemplarisch als Anzahl von Zitationen eines Werkes oder eines Autors bzw. als die Anzahl der Downloads eines Werkes oder die Zahl der Ausleihen im Bibliothekskontext operationalisiert sind. Ausgehend von diesen Popularitätsdaten wurden jene Merkmale als mögliche unabhängige Variablen in Betracht gezogen, die gegenwärtig in akademischen

[4] Die Anzahl der Stufen je UV werden miteinander multipliziert. Liegen beispielsweise 4 UV vor, von denen UV 1 und UV 2 jeweils zwei Stufen besitzen, während UV 3 auf drei und UV 4 auf vier Stufen variiert werden, handelt es sich um ein $2 \times 2 \times 3 \times 4$-Design mit insgesamt 48 experimentellen Bedingungen bzw. 48 Teilstichproben.

Suchsystemen integriert sind und zudem in Hinblick auf die Umsetzung des Designs eine möglichst homogene Art von Suchergebnissen zu erstellen erlauben, wie im Folgenden näher erläutert.

Im Gegensatz zu Zitationszahlen beziehen sich Downloadzahlen ausschließlich auf Dokumente in digitaler Form, während Ausleihzahlen im Bibliothekskontext sowohl von gedruckten Materialien als auch von E-Books ermittelt werden können. Allerdings ist eine direkte Vergleichbarkeit von Ausleihdaten mit Downloadzahlen aufgrund möglicher Verzerrungen nicht realisierbar, ohne die Daten mithilfe statistischer Verfahren, wie sie in der Bibliometrie verwendet werden, vorab zu normalisieren – eine Anforderung, die auch Zitationskennzahlen betrifft (Plassmeier et al., 2015).

Da den Kontext dieser Studie moderne akademische Suchsysteme (z. B. Google Scholar, ACM Digital Library) mit in die Suchergebnisdarstellung integrierten Popularitätsdaten wie die Zitations- und Downloadhäufigkeit eines Werkes bilden, wurden bibliothekarische Suchsysteme nicht berücksichtigt (vgl. Abschnitt 2.1.4). Vor diesem Hintergrund fiel die Entscheidung gegen die Untersuchung von Ausleihzahlen als unabhängige Variable und für die Beschränkung auf Downloadzahlen als Indikator für Nutzungshäufigkeit. Um eine weitgehend realistische Vergleichbarkeit von Zitationszahlen zu ermöglichen, wurden ausschließlich Werke aus Zeitschriftenaufsätzen als Surrogate erstellt (zur Problematik der Vergleichbarkeit der Surrogate vgl. Abschnitt 4.2.2).

Als unabhängige Variablen wurden schließlich die **Anzahl der Downloads eines Werkes (UV 1)**, die **Anzahl der Zitationen des Werkes (UV 2)** und die **Anzahl der Zitationen des Autors (UV 3)** festgelegt. Diese wurden jeweils auf den Stufen *gering* – *hoch* – *keine Angabe* manipuliert, d. h. die Kennzahl ist entweder gering oder hoch oder die Information (das Metadatum) ist nicht vorhanden (siehe Tabelle 4.2). Die UVn decken somit Indikatoren für Nutzungs- und Zitierhäufigkeiten ab und beziehen sich auf Kennzahlen, die sowohl das Werk als auch den Autor betreffen. Mit den drei UVn und jeweils drei Ausprägungen erhält das Design insgesamt 27 experimentelle Bedingungen, in denen alle möglichen Kombinationen aller UV-Stufen abgedeckt sind.

Tabelle 4.2 Anzahl und Stufen der unabhängigen Variablen

Nr.	UV	Stufe A	Stufe B	Stufe C
1	Anzahl Downloads	Geringe Anzahl	Hohe Anzahl	Keine Angabe
2	Anzahl Zitationen Werk	Geringe Anzahl	Hohe Anzahl	Keine Angabe
3	Anzahl Zitationen Autor	Geringe Anzahl	Hohe Anzahl	Keine Angabe

Anmerkung: Dieses $3 \times 3 \times 3$-Design ergibt 27 experimentelle Bedingungen.

Ein früheres Design sah als vierte UV den Impact eines Autors (z. B. h-Index) vor, die ebenfalls auf den Stufen *geringer Impact – hoher Impact – keine Angabe* manipuliert werden sollte. Der Pretest ergab jedoch, dass die aus den 81 resultierenden experimentellen Bedingungen und damit neun zu bearbeitenden Aufgaben mit jeweils neun zu bewertenden Surrogaten einen zeitlich unzumutbaren Bearbeitungsaufwand erfordert. Um die Bearbeitungsdauer von durchschnittlich 39 Minuten wesentlich zu verringen, wurde das Design auf drei UVn reduziert, wodurch anstelle der neun Aufgaben nur noch drei zu bearbeiten waren – eine Anzahl, die zufolge der Pretest-Erkenntnisse in Hinblick auf die Fragebogenseite mit den meisten Abbrüchen (bei Aufgabe 4)[5] die Motivation der Versuchspersonen scheinbar weitgehend ausschöpft.

Auf die Untersuchung eines möglichen Effekts des Autor-Impacts auf die AV wurde aus zwei Gründen verzichtet: (1) Bei dem Autor-Impact handelt es sich um eine Kennzahl, die im Gegensatz zu den anderen UVn nicht auf einer einfachen Summe beruht, sondern mittels anderer, intellektuell festgelegter Parameter errechnet wird und sich dahingehend von den anderen UVn unterscheidet. (2) Die Manipulation der einzelnen Stufen unterschied sich von der der anderen UVn dadurch, dass dem Surrogat im Pretest keine Zahlen für die Stufen *gering* und *hoch* zugewiesen waren, sondern ein prozentualer Wert, der die Zugehörigkeit einer Person zu einer Klasse innerhalb ihres wissenschaftlichen Feldes angibt (als geringer Impact wurde „Top 75 %“ oder „Top 50 %“, als hoher Impact „Top 5 %“ oder „Top 1 %“ festgesetzt). Dadurch wurden von den Versuchspersonen ein Umdenken und somit zusätzlicher kognitiver Aufwand während der Bearbeitung verlangt, denn eine kleine Zahl war in diesem Zusammenhang als höherer Impact, eine größere Zahl als geringerer Impact anzusehen.

[5] Aufgrund der randomisierten Reihenfolge der angezeigten Aufgaben ist für die höchste Abbruchquote bei Aufgabe 4 eine eher quantitative Ursache zu vermuten. Zu beachten ist, dass die Anzahl der Aufgaben in dem Pretest keine eigene zu untersuchende UV darstellt, da dies nicht der Zweck des Pretests war; auch auf die Berechnung der statistischen Signifikanz zwischen den Abbruchwerten wurde aus diesem Grund verzichtet.

4.1.1.1 Skalenniveau

Bei dem Skalenniveau der drei UVn handelt es sich um ein nominales bzw. kategoriales, auch wenn die Ausprägungen *gering* und *hoch* eine Rangordnung, also ein ordinales Skalenniveau, vermuten lassen. Wenn den Merkmalsausprägungen der UVn jeweils verschiedene (und immer andere) Zahlen zugeordnet werden, wie im hier beschriebenen Design, liegen ungeordnete Kategorien vor, die eine Nominalskala verlangen (Bortz & Schuster, 2010, S. 13). Tatsächlich verdeutlicht die dritte Stufe *keine Angabe* den kategorialen Charakter. Diese soll einerseits die Suchergebnisrepräsentation in traditionellen wissenschaftlichen Suchsystemen ohne integrierte Popularitätsdaten widerspiegeln und andererseits eine Baseline für eine eher thematische Relevanzbewertung darstellen. Diesbezüglich besteht die Annahme, dass eine Relevanzbewertung weitgehend auf Basis der thematischen Relevanz erfolgt, wenn die Versuchspersonen derjenigen experimentellen Bedingung ausgesetzt sind, in der alle drei UVn die dritte Ausprägung aufweisen, also die Angaben zu Zitations- oder Downloadzahlen in dem Surrogat fehlen.

4.1.1.2 Operationalisierung

Für die Zuweisung von Zahlen zu den Ausprägungen *gering* und *hoch* wurde aus forschungspragmatischen Gründen ein heuristischer Ansatz verfolgt: Für die UV-Stufen wurden Wertebereiche festgelegt, deren Zahlen von den Download- und Zitationszahlen in der internationalen informationswissenschaftlichen Community inspiriert sind.[6] Dazu wurde in Google Scholar nach viel und wenig zitierten Werken und Personen gesucht. Dabei dienten die Kennzahlen unter anderem von Diane Kelly oder Nick Belkin als weiche Bezugspunkte für die jeweils hohe Ausprägung, die Kennzahlen der Autorin und anderer informationswissenschaftlich Promovierender für die geringe Ausprägung. Zusätzlich wurde bei der Festlegung der Wertebereichsgrenzen darauf geachtet, eine Überschneidung zwischen den UVn zu vermeiden, um sicherzustellen, dass einzelne Zahlen nicht doppelt vorkommen bzw. eindeutig einer bestimmten UV-Stufe zuzuordnen sind. Außerdem kann eine intensivere Dosierung der UV, also eine stärkere Manipulation, einen positiven Einfluss auf die Teststärke haben (Döring & Bortz, 2016, S. 842). Dies motivierte zusätzlich, bei der Festlegung der Wertebereiche eher resolut als zu zaghaft vorzugehen.

[6] Studien aus der Bibliometrie zeigen, dass Zitierverhalten in verschiedenen wissenschaftlichen Fachdisziplinen unterschiedlich ist (Haustein, 2012, S. 223–299). Die Wahl der Informationswissenschaft als Fachdisziplin wurde im Zusammenhang mit den Merkmalen der Zielgruppe und der Wahl der Surrogate und Informationsbedürfnisse getroffen, wie unter Abschnitt 4.2 erläutert wird.

Die Wertebereiche (vgl. Tabelle 4.3) weisen bei UV 1 – Anzahl Downloads den größten Abstand auf, was die Dynamik dieser Kennzahl widerspiegelt. Als geringe Anzahl an Downloads wurden Zahlen von 210 bis 440, als hohe Zahlen von 5.400 bis 7.800 festgelegt. UV 2 – Anzahl Zitationen Werk bezieht sich wie UV 1 auf das einzelne Werk, stellt jedoch eine Kennzahl dar, die wesentlich langsamer wachsen kann, weil sie einen längeren Zeitraum benötigt (Kurtz & Bollen, 2010).

Die Anzahl der Zitationen eines Autors sind demzufolge in beiden Stufen der UV 3 höher als in denen der UV 2, weil diese alle zitierfähigen Werke betreffen, die die jeweilige Person bis dato publiziert hat. Anhand dieser Wertebereiche wurde mithilfe des kostenlos verfügbaren Online-Tools *Research Randomizer*[7] für alle Teilstichproben (Bedingungen) jeweils eine einmalig vorkommende Zufallszahl ausgewählt (vgl. Tabelle 4.4). Da die Versuchspersonen allen 27 experimentellen Bedingungen ausgesetzt werden sollten und dies der Anzahl der insgesamt zu bewertenden Surrogate entspricht, waren für alle UVn jeweils 9 Werte erforderlich. Ein Surrogat weist in einer bestimmten Bedingungsvariation für jede Stufe einen einmaligen Wert auf, der sich in derselben Stufe einer anderen Bedingung nicht wiederholt. Liegt zum Beispiel für das erste Surrogat die Kombination UV 1 Stufe 1 – UV 2 Stufe 2 – UV 3 Stufe 3 vor, erhält das Surrogat die Werte 240 – 106 – keine Angabe. Die Wertevergabe zu den Surrogaten ergibt sich aus dem faktoriellen Design, in dem die Zuweisung der Surrogate zu den experimentellen Bedingungen in randomisierter Reihenfolge erfolgte (vgl. Abschnitt 4.1.5).

Eine alternative Möglichkeit der Wertezuweisung zu den UV-Stufen A und B läge in einer Ad hoc-Generierung aus den zuvor festgelegten Wertebereichen für die jeweilige Bedingung, also das jeweilige Surrogat, während der Bearbeitung des Online-Fragebogens. Das heißt, dass in dem Fall für eine Versuchsperson kein Surrogat dasselbe wie für eine andere innerhalb derselben Bedingung wäre. Dieser Umstand hätte eine zusätzliche Variation im Experiment bedeutet, die für die spätere statistische Analyse als wenig zielführend erachtet und daher abgelehnt wurde.[8]

[7] https://www.randomizer.org/ (letzter Zugriff: 03.04.2020).

[8] Generell ist zu berücksichtigen, dass eine Ad hoc-Generierung ggf. eine Anpassung der Syntax in der Software, mit dem der Fragebogen erstellt und die Befragung online durchgeführt wird, und einen entsprechenden Programmieraufwand in Verbindung mit Tests erfordert; aus forschungspragmatischen Gründen sollte diesbezüglich ein Abwägen des absehbaren Aufwands gegenüber dem erhofften Mehrwert erfolgen.

Tabelle 4.3 Wertebereiche der UV-Stufen A und B

	Geringe Ausprägung		Hohe Ausprägung	
UV	Von	bis	von	bis
Anzahl Downloads	210	440	5.400	7.800
Anzahl Zitationen Werk	2	9	70	160
Anzahl Zitationen Autor	470	650	2.500	4.600

Tabelle 4.4 Werte der UV-Stufen A und B

UV 1 – Downloads			UV 2 – Zit. Werk			UV 3 – Zit. Autor		
Ausprägung			Ausprägung			Ausprägung		
#	Gering	Hoch	#	Gering	Hoch	#	Gering	Hoch
1	240	5.490	1	9	106	1	603	2.975
2	294	5.930	2	6	97	2	638	3.985
3	384	7.125	3	2	102	3	502	3.241
4	423	5.988	4	5	112	4	552	2.950
5	359	6.400	5	4	95	5	472	4.302
6	224	7.565	6	3	149	6	551	3.447
7	433	6.606	7	7	136	7	497	2.688
8	389	6.839	8	9	83	8	525	4.070
9	233	7.030	9	2	71	9	576	3.753

4.1.2 Abhängige Variable

Zur Erforschung von Relevanzkriterien gibt es zahlreiche Studien, in denen weder implizite Relevanzbewertungen erfasst noch explizite Relevanzbewertungen von menschlichen Juroren erhoben wurden; stattdessen führten die Forschenden in solchen Studien Interviews durch oder wendeten die Methode des lauten Denkens an (vgl. Abschnitt 2.2.1).

Mit der Forschungsfrage F2 nach dem Effekt von Popularitätsdaten auf die Relevanzbewertung ist bereits festgelegt, dass Relevanzbewertungen zu erheben sind. Implizite Relevanzbewertungen, die in IR-Studien auf Basis von Klicks und/oder Verweildauer abgeleitet wurden, waren als mögliche AV in dieser Studie bereits zu einem frühen Zeitpunkt ausgeschlossen. Klicks sind aufgrund

ihrer lediglich binären Ausprägung nicht mit Relevanzbewertungen gleichzusetzen (vgl. Abschnitt 2.2.2); die Verweildauer als Relevanzbewertung zu betrachten ist ebenso wenig angemessen, weil diese Kennzahl nur im Zusammenhang mit tatsächlichen Bewertungen Einblicke in das Bewertungsverhalten bieten kann (Kelly & Belkin, 2004).

Die abhängige Variable ist die explizite Relevanzbewertung der Suchergebnisse, die anhand der von den Versuchspersonen vorgenommenen Kennzeichnungen auf einer Skala erfasst wurden. Diese Skala war mit den Polen 0 (links) und 100 (rechts) versehen, d. h. die AV weist 101 Merkmalsausprägungen auf, wobei die Punkte auf der Skala für die Versuchspersonen nicht sichtbar waren, ebenso wenig die Messwerte der einzelnen Beurteilungen. Diese hohe Zahl an Skalenstufen (Rating-Kategorien) ist nicht nur mit dem nicht-binären Verständnis von Relevanz zu begründen, vielmehr soll auf diese Weise dem Granularitätsanspruch der *graded relevance* Rechnung getragen werden (Roitero et al., 2018). Der nachfolgende Abschnitt 4.1.2.1 widmet sich den Erkenntnissen bisheriger Studien zur Erhebung von Bewertungen, die *graded relevance* berücksichtigen und begründet die Wahl der für das hier beschriebene Experiment verwendeten Skalenart. Daran anschließend werden das Skalenniveau und die ausgewählte Skala vorgestellt (Abschnitt 4.1.2.2), bevor die Operationalisierung der abhängigen Variablen erörtert wird (Abschnitt 4.1.2.3).

4.1.2.1 Erhebung von Graded Relevance Assessments

Traditionell wurden in der Information Retrieval-Evaluierung wie in klassischen TREC-Studien ausschließlich binäre Bewertungen erhoben – ein Dokument war entweder relevant oder nicht, was auf die auf dem Cranfield-Paradigma beruhenden, stark vereinfachten Annahmen über die Bewertung von Dokumenten in einer Testkollektion zurückzuführen ist (Buckley & Voorhees, 2005). Da eine dichotome Nominalskala lediglich zwei Kategorien besitzt, wird eine hohe Toleranz bei den Bewertungen von Dokumenten in Kauf genommen. Mehrere Abstufungen bei der Relevanzbewertung sind jedoch sinnvoll zur Unterscheidung von Rankingalgorithmen, um diejenigen Verfahren, die in der Lage sind hochrelevante Ergebnisse auf den höheren Positionen einer Trefferliste zu produzieren, von denjenigen, die nur wenig relevante Ergebnisse erzeugen können, zu unterscheiden (Kekäläinen & Järvelin, 2002). Vor diesem Hintergrund stellt sich die Frage, welche Skalenart bzw. welche Anzahl an Stufen einer Skala am besten zur Erhebung von Relevanzbewertungen geeignet ist.

Studien, die der Frage nachgingen, wie Relevanzbewertungen (im Forschungskontext) am besten Ausdruck verliehen werden kann, wurden bereits in den

frühen Jahren der Relevanzforschung durchgeführt (vgl. Mizzaro, 1997). Hervorzuheben sind hier die Arbeiten des Teams um Cuadra und Katter von der damaligen System Development Corporation, das mehrere Experimente zu Relevanzbewertungen und deren Einflussfaktoren durchführte und veröffentlichte (Cuadra & Katter, 1967a, 1967b; Katter, 1968; System Development Corporation, 1967). Die Ergebnisse der Experimente zeigten, dass Relevanzbewertungen unter anderem durch die Art der verwendeten Skala beeinflusst werden (Cuadra & Katter, 1967b). Beispielsweise berichtet Katter (1968) von einem Vergleich zwischen Relevanzbewertungen mittels neunstufiger Ratingskala und Bewertungen mithilfe von paarweisen Dokumenten-Rankings. Die Ergebnisse zeigen eine Verzerrung bei der Bewertung mittels Rankings, die Katter als „cascaded distortion process" bezeichnet: Während beim Rating ein Wert mehrmals vergeben wird, ist dies beim Ranking nicht möglich, da jede Position nur einmal besetzt werden kann. Solche Einschränkungen können zu Verletzungen der intendierten Relevanzbewertungen führen. Zudem wurde geschlussfolgert, dass die Verwendung einer Ratingskala mit 6 bis 8 Kategorien effektiver war als die von Skalen mit weniger Kategorien (Cuadra & Katter, 1967b). Ähnliches stellten Tang, Shaw, & Vevea (1999) bei einem Vergleich von Skalen mit 2 bis 11 Punkten fest: Sie identifizierten die siebenstufige Skala als diejenige, mit der die Studienteilnehmenden das höchste Maß an Selbstbewusstsein („the highest level of confidence") bezüglich ihrer Bewertungen zum Ausdruck brachten.

Nichtsdestotrotz werden in Information Retrieval-Studien häufig Ordinalskalen verwendet, die lediglich drei oder ähnlich einer klassischen Likert-Skala fünf Stufen aufweisen (Kelly, 2009, S. 41, 42). Obwohl anhand dieser Abstufungen zwar Aussagen darüber getroffen werden können, dass ein Dokument wesentlich relevanter ist als ein anderes, ist es nicht zulässig, Aussagen über die Differenzen zu treffen. So kann beispielsweise nicht gefolgert werden, dass bei einer 5-Punkte-Skala (nicht relevant – wenig relevant – teilweise relevant – ziemlich relevant – hoch relevant) ein Dokument mit dem Messwert 4 doppelt so relevant ist wie ein Dokument mit dem Messwert 2. Die Stufen auf einer solchen Skala sind demzufolge als Kategorien zu verstehen und bilden dennoch keine Nominalskala, da sie eine Ordnung erlauben. (Kelly, 2009, S. 42)

Eine Form der metrischen Skala, die in jüngeren IR-Studien zur Erhebung von Relevanzbewertungen verwendet wurde, ist die als Verhältnisskala (auch Ratioskala) entwickelte *Magnitude Estimation* (ME)[9], die eine feinstufigere Bewertung erlaubt als eine dichotome Skala oder Ordinalskala (Maddalena et al., 2017;

[9] In der Psychophysik zur numerischen Erfassung von subjektiven Intensitäten durch Stimuli entwickelt, wird ME heutzutage unter anderem auch in der Usability-Forschung eingesetzt,

Roitero et al., 2018; Turpin et al., 2015). Mit der ME bewerten Jurorinnen und Juroren den subjektiv wahrgenommenen Relevanzwert eines Dokuments in Relation zum jeweils zuvor gesehenen Dokument, ausgehend von einem selbstgewählten Startwert und ohne feste Wertebereichsgrenzen (Turpin et al., 2015). Wenn zum Beispiel ein Dokument 1000 Punkte erhält und das folgende als halb so relevant erachtet wird, vergibt die Person 500 Punkte; ist das darauffolgende wieder wesentlich relevanter, könnte die Person 850 Punkte vergeben. Das bedeutet, dass jeder Juror die Nutzung der ME individuell an die eigenen Präferenzen bzgl. einer Skala anpassen kann. Ein Vorteil dieser Punktevergabe ist, dass den Juroren „niemals die Werte ausgehen", da immer ein noch kleinerer oder größerer Wert an ein Dokument vergeben werden kann; von Nachteil ist allerdings, dass die Bewertungen zwischen Juroren und Aufgaben ohne eine Normalisierung der Werte nicht miteinander verglichen werden können, und dass diese etwas unnatürliche Form der Bewertung eine gewisse Gewöhnungsphase erfordert (Roitero et al., 2018).

Eine andere Einsatzmöglichkeit der ME erfolgte dagegen in früheren Studien (Bruce, 1994; Eisenberg, 1988; Eisenberg & Hu, 1987; Janes, 1991a), in denen Probanden auf einer 100 Millimeter langen Linie auf einem Blatt Papier ein Kreuz setzten, das den Messwert markiert. Eisenberg (1988) kam zu dem Ergebnis, dass der Einsatz einer solchen ME-Skala mit 101 Abstufungen zur Erfassung von Relevanzbewertungen geeignet und empfehlenswert sei.

Dieselbe Methode wurde verwendet zur Erforschung des *break point*, also des Punktes, ab dem für die Jurorinnen und Juroren ein Dokument als relevant gilt, im Sinne der Frage „Wie relevant muss ein Dokument sein, damit es als relevant bewertet wird?" Ergebnisse zweier Studien zeigen, dass bei 41 von 100 Bewertungspunkten (Eisenberg & Hu, 1987) bzw. bei 46 von 100 Punkten (Janes, 1991b) diese Schwelle erreicht ist. Nicht nur Relevanzbewertungen wurden mittels ME erhoben: Bruce (1994) ließ Probanden auf diese Weise Dokumenteigenschaften und Informationsattribute nach ihrer Wichtigkeit bezüglich der wahrgenommenen Relevanz bewerten und resümierte, dass die Methode einen beobachtbaren Effekt des Einflusses der Interaktion mit dem IRS auf die nutzerzentrierte Relevanzbewertung erlaubt.

Neuere Erkenntnisse stützen die Ergebnisse der früheren Studien. So verglichen Roitero u. a. (2018) Relevanzbewertungen, die mittels vier verschiedener Skalen erhoben wurden: dichotome Skala (binäre Sicht auf Relevanz), 4-stufige Ordinalskala, *Magnitude Estimation* in Form von Relationsbewertungen, und

---

um in Anschlussfragebögen zu erfassen, wie die Teilnehmenden die Schwierigkeitsgrade der gestellten Aufgaben beurteilten (Sauro & Lewis, 2012).

S100, einer *fine-grained relevance scale* mit 101 Abstufungen, die der Einsatzform von ME als Liniendarstellung entspricht. Die Autoren stellten fest, dass S100 effektiv, robust und brauchbar ist zur Erhebung von feingranulierten Relevanzurteilen.

Zusammengefasst folgt die Auswahl der Skala für die hier beschriebene Studie den Erkenntnissen und Forderungen der frühen Studien zum Einsatz von mehrstufigen Skalen und zugleich den neueren Erkenntnissen in Hinblick auf S100, mit der die Testpersonen feinstufige Relevanzbewertungen vornehmen können.

4.1.2.2 Skalenniveau und Skalenauswahl

Das Skalenniveau in der hier beschriebenen Studie ist metrisch; es liegt eine Intervallskala vor, die eine lineare Transformation und Aussagen über die Gleichheit von Differenzen zulässt im Gegensatz zu Ordinal- und Nominalskalen (Döring & Bortz, 2016, S. 233). Letzteres ist insbesondere vor dem Hintergrund des gewählten Within-Subjects-Designs ein weiteres Argument für die Wahl einer Intervallskala, um Differenzen der Bewertungen innerhalb der Versuchspersonen zwischen den Bedingungen berechnen und auch kleine Effekte einer UV aufdecken zu können.

Konkret handelt es sich in der Umsetzung der Intervallskala um eine Kombination aus visueller Analogskala (*visual analogue scales,* VAS) und Schiebereglerskala (*slider scale*) – zwei Skalen, die sich prinzipiell sehr ähnlich sind: Während die Versuchspersonen bei einer visuellen Analogskala nach der *point & click*-Methode vorgehen, erfolgt die Beurteilung auf einer Schiebereglerskala mittels *drag & drop* des Schiebereglers (Funke, 2016). Die VAS ist eine Variante der Ratingskala, bei der Versuchspersonen ihre Urteile über Merkmalsausprägungen ohne Abschnitts- bzw. Stufenmarkierungen mit dem Setzen eines Kreuzes bzw. in der computergestützten Befragung mittels Radio Button angeben (Döring & Bortz, 2016, S. 246 ff.). Die Position des Kreuzes bzw. Radio Buttons entspricht dem auf der Skala liegenden Messwert. Dasselbe trifft auch auf die Schiebereglerskala zu (Funke, 2016).

Der Vorteil der Schiebereglerskala kann in einer höheren Sensitivität gesehen werden: Durch das stufenlose Schieben des Reglers zum linken oder rechten Pol hin scheint es leichter, auch feinere Abstände von beispielsweise unter 10 Punkten zwischen bereits gesetzten Bewertungen per *drag & drop* (oder mittels Pfeiltasten) zu erzeugen und auf diese Weise auch geringe Urteilsdifferenzen (innerhalb der Versuchsperson aufgrund des Within-Subjects-Designs) anzuzeigen als durch erneutes *point & click*. Nachteilig ist allerdings die höhere Anzahl an Aktionen, die die Versuchsperson beim Setzen des Schiebereglers (Mauszeiger bewegen – Maustaste klicken und halten – Mauszeiger bewegen – Maustaste

loslassen) im Vergleich zum Setzen des Wertes bei einer VAS (Mauszeiger bewegen – Maustaste klicken) durchzuführen hat (Funke, 2016, S. 245, 246). Dieser etwas höhere Aufwand erschien allerdings vor dem Hintergrund der als ohnehin anspruchsvoll anzusehenden Bewertungsaufgaben (Informationsbedürfnisse, vgl. Abschnitt 4.2.1) für den Forschungszweck vertretbar.

Ferner kann der Umgang mit einem Schieberegler zu negativen Effekten führen (Toepoel & Funke, 2018), wie z. B. zu einer höheren Abbruchrate und einem höheren Zeitaufwand, wobei Schwierigkeiten im Umgang mit Schiebereglern von dem Bildungsgrad der Teilnehmenden abhängig sind (Funke et al., 2011). So ist davon auszugehen, dass die Versuchspersonen der Zielgruppe in der hier beschriebenen Studie, also Personen mit mindestens einem akademischen Abschluss auf Master-Niveau und in der Wissenschaft tätig (vgl. Abschnitt 4.2), über die erforderlichen kognitiven Fähigkeiten zum korrekten Verständnis der Funktionsweise eines Schiebereglers verfügen.[10]

Ein gravierender Nachteil der Schiebereglerskala besteht in der Vorabpositionierung des Schiebereglers durch die Forschungsleitung, wodurch zum einen eine mögliche Beeinflussung der Versuchspersonen durch die Ausgangsposition nicht auszuschließen ist, zum anderen eine Interpretation des Messwerts schwierig wird, wenn dieser mit dem Wert der Startposition übereinstimmt (Döring & Bortz, 2016, S. 248). Bei der hier berichten Studie wurde diesem Problem dadurch begegnet, dass der Schieberegler als zunächst nicht sichtbar eingestellt war und erst durch Mausklick auf einen Punkt der Skalenlinie erschien (vgl. Abschnitt 4.2.3.3), was dem klassischen VAS-Ansatz entspricht (*point & click*). Auf diese Weise entschied die Versuchsperson eigenständig über die Ausgangsposition des Schiebereglers.

Weiterhin sollte es möglich sein, eine Bewertung nachträglich zu ändern, also die Position des bereits gesetzten Schiebereglers anzupassen. Somit ist sichergestellt, dass die Relevanzbewertung einzelner Suchergebnisse unter Berücksichtigung aller Suchergebnisse in der jeweiligen Aufgabe vorgenommen werden kann. Der Entscheidung für eine nachträgliche Urteilsanpassung liegt die Annahme zugrunde, dass die Relevanzbewertung eines Dokuments nicht unabhängig von der der anderen Dokumente innerhalb eines Korpus bzw. die Bewertung eines Suchergebnisses nicht losgelöst von der der anderen Ergebnisse in der Trefferliste erfolgt (vgl. Abschnitt 4.1.5.2). Eine erneute Bewertungsmöglichkeit bietet

[10] Damit Versuchspersonen die Funktionsweise der Skala im Fragebogen verstehen, sollte während der Befragung und vor der zu bearbeitenden Aufgabe ein Beispiel gezeigt und ggf. anhand einer Testskala den Teilnehmenden die Möglichkeit zum Ausprobieren gegeben werden. Im Rahmen dieser Studie wurde dies im Fragebogen entsprechend umgesetzt (vgl. Abschnitt 4.2.3.3).

zwar nicht nur der Einsatz eines Schiebereglers, weil diese Einstellung je nach Fragebogensoftware auch bei anderen Messinstrumenten im Rahmen der Fragebogenkonstruktion vorgenommen werden kann; ein Schieberegler impliziert diese Möglichkeit jedoch unmittelbar, ohne dass in den Aufgabeninstruktionen ausdrücklich darauf hinzuweisen wäre.

4.1.2.3 Operationalisierung von Relevanz als Nützlichkeit

Nützlichkeit, also *usefulness*, wurde von verschiedenen Verfechtern der nutzerzentrierten Relevanzperspektive als angemesseneres Konzept für die Evaluierung von Information Retrieval-Systemen im Gegensatz zu einer rein systembasierten Perspektive auf Relevanz als eine auf Aboutness abzielende Bewertung vorgeschlagen (vgl. Abschnitt 2.1.1 und Abschnitt 3.1.1). Es wurde erläutert, dass relevante Informationen nicht nur thematisch passend zur Suchanfrage oder zum Informationsbedürfnis, sondern auch nützlich zur Befriedigung des Informationsbedürfnisses und darüber hinaus nützlich zur Lösung des Problems bzw. der Bewältigung der Aufgabe sein müssen (Mizzaro, 1997; Saracevic, 2016b).

In Studien, in denen Jurorinnen und Juroren Dokumente in Hinblick auf deren Nützlichkeit oder Nutzen bewerteten, geschah dies im Einklang mit einer ziel- und aufgabenorientierten Definition von Relevanz. So verfolgten Cool, Belkin, Kantor, & Frieder (1993) das Ziel, unter anderem die folgenden Forschungsfragen zu beantworten:

> What are the relationships between a person's goals (or information problems) and the documents used in responding to those goals (problems)? That is, what are the uses that people will make of the documents, and how do they judge (evaluate) documents with respect to those uses? Are there characteristics other than topical relevance which affect a person's evaluation of a document's usefulness? (Cool et al., 1993, S. 77)

In einer Studie zu dem Einfluss der Reihenfolge der zu bewertenden Dokumente operationalisierten Xu & Wang (2008) Relevanz ebenfalls als Nützlichkeit in Hinblick auf den aufgabenorientierten Charakter des Informationssuchprozesses:

> All returned documents were evaluated based on participants' perception of relevance. The term "usefulness" was used in place of "relevance" to make the concept more straightforward to the participants […] and corresponds better to the definition of situational relevance in a task-oriented search. (Xu & Wang, 2008, S. 1269)

Im Unterschied zur erstgenannten Studie wurden bei Xu & Wang (2008) die Bewertungen auf Basis von Webseiteninhalten getroffen, also *evaluative judgments* vorgenommen, während bei Cool, Belkin, Kantor, & Frieder (1993)

predictive judgments erbeten wurden, was sich in der Instruktionsformulierung an die Teilnehmenden niederschlägt, denn die Aufgabe bestand darin, „[to] indicate whether they *thought* that they *would* use [the document] for their essay" (S. 78; Kursivdruck im Original nicht enthalten). Analog zum Relevanzkonzept stellt sich die Frage, ob die Nützlichkeit eines Dokuments zum Zeitpunkt der Sichtung der Surrogate oder zu einem wesentlich späteren Zeitpunkt, der weit nach dem Abschluss der interaktiven Suche im IR-System liegt, und nach der tatsächlichen Nutzung des Dokuments beurteilt werden kann.[11] Letzteres wäre auch unabhängig von einer subjektiven Bewertung feststellbar:

> Usefulness or utility could be determined subjectively by the user or objectively by looking at whether or not the user used the document, the contact time with the document, or the results of contact with the document, such as improved productivity, development of a new product, or publication. (Tague-Sutcliffe, 1992, S. 474–475)

Die aufgaben- und zielorientierte Definition von Relevanz, die im Rahmen der Konzeptspezifikation dieser Arbeit zugrunde gelegt wird (vgl. Abschnitt 3.1.1), findet sich in der Formulierung der Bewertungsaufforderung und der Beschriftung der beiden Skalenenden wieder. So wurde auf die Verwendung des Relevanzbegriffs gezielt verzichtet und Relevanz als Nützlichkeit (*usefulness*) operationalisiert: Der linke Pol der Skala ist als „überhaupt nicht nützlich", der rechte Pol der Skala als „sehr nützlich" gekennzeichnet. Die Versuchspersonen sollten beurteilen, für wie nützlich sie jedes Suchergebnis zur Befriedigung des zuvor gelesenen Informationsbedürfnisses halten. Bei den erhobenen Bewertungen handelt es sich daher um *predictive judgments*, die die vermutete Nützlichkeit im Kontext des Informationsbedürfnisses betreffen und sich auf die Relevanzbewertung des Surrogats beschränken (vgl. Abschnitt 2.1.4).

4.1.3 Hypothesen

Die zu prüfenden Hypothesen stellen Erwartungen hinsichtlich der Beantwortung der Forschungsfrage F2 (*Welchen Einfluss haben Popularitätsdaten auf die Bewertung der Relevanz von Suchergebnissen in akademischen Suchsystemen?*) dar. Auf

[11] Hieran zeigt sich erneut die Wichtigkeit der Unterscheidung von Bewertungen in *predictive judgments* und *evaluative judgments*, möglicherweise weniger in Hinblick auf die Wahl des Konzepts oder der Operationalisierung als vielmehr vor dem Hintergrund, auf Basis von Relevanzbewertungen aussagekräftige Rückschlüsse auf interne Bewertungsmodelle ziehen und dadurch Relevanz besser verstehen zu können.

der Basis der Erkenntnisse aus der Literaturschau wird davon ausgegangen, dass Popularitätsdaten einen positiven Einfluss auf die Relevanzbewertung haben, was sich in einer höheren Bewertungspunktzahl zeigt.

Diese Annahme wird zum einen mit der Erkenntnis aus der Arbeit von Rieh (2002) begründet, dass anhand zusätzlicher Informationen zu einem Suchergebnis informationssuchende Personen bessere *predictive judgments* hervorbringen (vgl. Abschnitt 2.1.2.1, S. 30), wobei sich die Güte in der größtmöglichen Übereinstimmung mit dem *evaluative judgment* ausdrückt, welches in dieser Arbeit jedoch nicht überprüft wird. Zum anderen deckte bereits Wang (1994) im Rahmen ihrer Studie zur Dokumentenauswahl den Bedarf von informationssuchenden Personen auf, Informationen über den Autor eines Werks während der Interaktion mit einem akademischen Suchsystem in die Entscheidungsfindung zur Dokumentenauswahl miteinzubeziehen, welcher auf den besonderen Stellenwert des Kriteriums Autorität hindeutet (vgl. Abschnitt 2.1.4, S. 41).

Für das vorliegende Experiment lassen sich drei Hypothesen über die Haupteffekte, also die Wirkungen der einzelnen unabhängigen Variablen (UVn) unabhängig von den Stufen der jeweils anderen UV (Sedlmeier & Renkewitz, 2018, S. 172), aufstellen:

H1: **Die Downloadhäufigkeit eines Werks hat einen positiven Einfluss auf die Relevanzbewertung.** Bei einer hohen Zahl von Downloads eines Werkes ist die Punktzahl der Relevanzbewertung im Durchschnitt größer als bei einer geringen Anzahl oder bei Nichtanzeige (k.A.); bei Nichtanzeige ist die Punktzahl der Relevanzbewertung im Durchschnitt kleiner als bei einer geringen oder hohen Anzahl Downloads.

H2: **Die Zitationshäufigkeit eines Werkes hat einen positiven Einfluss auf die Relevanzbewertung.** Bei einer hohen Anzahl von Zitationen eines Werkes ist die Punktzahl der Relevanzbewertung im Durchschnitt größer als bei einer geringen Anzahl oder bei Nichtanzeige (k.A.); bei Nichtanzeige ist die Punktzahl der Relevanzbewertung im Durchschnitt kleiner als bei einer geringen oder hohen Anzahl von Zitationen.

H3: **Die Zitationshäufigkeit des Autors hat einen positiven Einfluss auf die Relevanzbewertung.** Bei einer hohen Anzahl von Zitationen eines Autors ist die Punktzahl der Relevanzbewertung im Durchschnitt größer als bei einer geringen Anzahl oder bei Nichtanzeige (k.A.); bei Nichtanzeige ist die Punktzahl der Relevanzbewertung im Durchschnitt kleiner als bei einer geringen oder hohen Anzahl von Zitationen eines Autors.

Die Annahmen lassen sich anhand der beiden experimentellen Bedingungen, in denen alle drei Stufen jeweils am weitesten auseinander liegen, konkreter beschreiben. In der Bedingung, in der alle drei UVn auf der dritten Stufe (keine Angabe) manipuliert sind, dürften die Bewertungen der Probanden für dieses Surrogat im Durchschnitt die geringsten Punktzahlen aufweisen; im Gegensatz dazu werden in der Bedingung, in der alle UVn die Ausprägung mit der hohen Anzahl besitzen, die Bewertungen für dieses Surrogat im Durchschnitt vermutlich die größten Punktzahlen zeigen.

Das mehrfaktorielle Untersuchungsdesign erlaubt es zudem zu prüfen, ob Interaktionseffekte vorliegen, d. h. ob die Wirkung einer UV abhängig ist von der Ausprägung einer anderen UV (Sedlmeier & Renkewitz, 2018, S. 171). Dies trägt insbesondere zu der Beantwortung der Forschungsfrage F3 (Welche Popularitätsdaten beeinflussen die Relevanzbewertung in welchem Maße?) bei. Möglich sind im vorliegenden Experiment Interaktionseffekte der 1. Ordnung zwischen UV1 und UV2, UV1 und UV3 sowie UV2 und UV3, und ein Interaktionseffekt der 2. Ordnung, also zwischen UV1 und UV2 und UV3. Dazu werden die Effekte ermittelt, die zum Beispiel UV 1 auf allen Stufen von UV 2 aufweist, und miteinander verglichen. Sind die Effekte von UV 1 auf den jeweiligen Stufen von UV 2 ungleich groß, liegt eine Interaktion vor; sind die Effekte gleich groß, liegt keine Interaktion vor (Sedlmeier & Renkewitz, 2018, S. 172–173). Da zur Hypothesenaufstellung derartiger Interaktionseffekte Erkenntnisse aus sehr ähnlichen experimentellen Untersuchungen über den Einfluss bestimmter Merkmale von Surrogaten auf die Relevanzbewertung herangezogen werden müssten, solche jedoch nicht vorliegen, lassen sich diesbezüglich keine Hypothesen formulieren.

4.1.4 Umgang mit Störvariablen

Neben den drei unabhängigen Variablen gibt es möglicherweise weitere Variablen, die einen Einfluss auf die abhängige Variable ausüben. Solche Drittvariablen[12] können zum einen Merkmale sein, die im Zuge des Experiments ebenfalls erhoben, aber nicht durch die Forschungsleitung manipuliert werden, wie beispielsweise soziodemografische Merkmale der Teilnehmenden (Kelly,

[12] Die American Psychological Association bezeichnet diese Drittvariablen als *quasi-independent variables* (https://dictionary.apa.org/quasi-independent-variable, letzter Zugriff: 09.04.2020) ebenso wie Kelly (2009); die gängigen deutschsprachigen Methodenlehrbücher (Bortz & Schuster, 2010; Döring & Bortz, 2016; Eid et al., 2017; Sedlmeier & Renkewitz, 2018) enthalten den Begriff *quasi-unabhängige Variable* nicht.

2009, S. 38). Zum anderen können potenziell konfundierende Variablen (Störvariablen) auftreten, die es, um der internen Validität und schließlich der Güte des Experiments willen, weitestgehend zu kontrollieren gilt. Das Ziel der Kontrolle möglicher Störvariablen besteht darin, Alternativerklärungen des beobachteten Effekts ausschließen und somit kausale Schlussfolgerungen ableiten zu können (vgl. Abschnitt 2.2.4).

Neben den üblicherweise erhobenen soziodemografischen Merkmalen waren für die hier beschriebene Studie weitere Drittvariablen von Interesse. Dazu zählten unter anderem der bisherige höchste Bildungsabschluss, die wissenschaftliche Fachdisziplin der Versuchspersonen und Informationen über Erfahrungen mit wissenschaftlichen Suchsystemen. Auf diese Drittvariablen wird in Abschnitt 4.2.3 im Zusammenhang mit der Fragebogenentwicklung näher eingegangen.

Mögliche Störvariablen bestehen hinsichtlich der Untersuchungspersonen (personengebundene Störvariablen), der Versuchssituation (z. B. Umgebungseinflüsse) und der Versuchsleitung (Erwartungseffekte) (Sedlmeier & Renkewitz, 2018, S. 139 ff.). Wie oben erwähnt besteht ein Vorteil des Within-Subjects-Designs darin, dass alle personengebundenen Störvariablen „perfekt parallelisiert" (Sedlmeier & Renkewitz, 2018, S. 157) sind. Somit ist das Problem personengebundener Störvariablen für diese Studie gelöst.

Externe Störeinflüsse auf die Versuchssituation wie Ablenkungen durch Lärm oder Unterbrechungen der Untersuchungspersonen können durch die Versuchsleitung nicht eliminiert werden, da das Online-Experiment im Feld stattfindet. Ein Konstanthalten möglicher Störeinflüsse erfolgt jedoch dadurch, dass alle Teilnehmenden dieselben Instruktionen in allen Bedingungen erhalten und alle dasselbe Interface verwenden, d. h. für alle erfolgt der Versuchsablauf unter identischen Bedingungen. Identische Bedingungen bedeuten jedoch nicht, dass die Versuchspersonen die Aufgaben in identischer Reihenfolge bearbeiten. Stattdessen werden sowohl die Reihenfolge der Aufgaben als auch die Reihenfolge der angezeigten Surrogate während des Online-Experiments ad hoc randomisiert. Auf diese Weise wird vermieden, dass Lerneffekte und Positionseffekte einen unerwünschten Einfluss auf die Ergebnisse bewirken. Generell ist damit zu rechnen, dass die Teilnehmenden während des Experiments Vermutungen über Hypothesen oder erwünschte Ergebnisse anstellen. Merkmale der Instruktionen oder des Stimulusmaterials können bestimmte Hinweise auf den eigentlichen Zweck der Untersuchung geben (sog. *demand characteristics*), was manche Teilnehmenden

dazu veranlassen kann, sich gezielt erwartungskonform zu verhalten[13] (Sedlmeier & Renkewitz, 2018, S. 147). Solche Erwartungseffekte sollen durch das Durchführen eines Experiments als Blindversuch kontrolliert werden. Aus diesem Grund wurden die Teilnehmenden erst nach Ende des Experiments über dessen Zweck und vorgenommene Manipulationen aufgeklärt (vgl. Abschnitt 4.2.3).

In der Versuchsleitung können ebenfalls Erwartungseffekte die Ergebnisse des Experiments beeinflussen. Da es sich jedoch um ein Online-Experiment handelt, also das Experiment nicht in einem Labor stattfindet, ist die Versuchsleitung lediglich durch die Instruktionen im Fragebogen bzw. durch die Informationen in den Einladungen zur Teilnahme per E-Mail repräsentiert, aber nicht physisch anwesend. Eine individuelle Interaktion zwischen Versuchsleitung und Versuchsperson findet nicht statt, wodurch eine unerwünschte Beeinflussung der Versuchsperson durch die Versuchsleitung (und umgekehrt) ausgeschlossen werden kann.

4.1.5 Versuchsaufbau

Das mehrfaktorielle Within-Subjects-Design wurde als vollständiger Versuchsplan umgesetzt, in dem alle insgesamt 27 möglichen Kombinationen der Stufen jeder unabhängigen Variablen enthalten sind. Auf diese Weise wird einer möglichen Konfundierung von Versuchsbedingungen begegnet und die interne Validität des Experiments erhöht (Sedlmeier & Renkewitz, 2018, S. 138 ff.). Tabelle 4.5 zeigt den vollständigen faktoriellen Versuchsplan mit allen 27 Kombinationen aller unabhängigen Variablen auf allen Stufen.

In Abbildung 4.1 ist der faktorielle Versuchsplan formalisiert in der weit verbreiteten Notation von Shadish, Cook, & Campbell (2002) dargestellt. Die Notation erhält die folgenden drei Elemente (Döring & Bortz, 2016, S. 102; Shadish u. a., 2002, S. 263):

X = Treatmentbedingung/unabhängige Variable;
O = Beobachtung/Messung/abhängige Variable und
R = Randomisierung

[13] Dieses Verhalten wird in englischsprachigen Quellen als *good subject effect* bezeichnet, der im Zusammenhang mit der Erhebung von Relevanzbewertungen erneut aufgegriffen wird (vgl. Abschnitt 4.1.5).

Tabelle 4.5 Vollständiger faktorieller Versuchsplan

	A_1						A_2						A_3					
	B_1		B_2		B_3		B_1		B_2		B_3		B_1		B_2		B_3	
C_1	1	S_{111}	2	S_{121}	3	S_{131}	4	S_{211}	5	S_{221}	6	S_{231}	7	S_{311}	8	S_{321}	9	S_{331}
C_2	10	S_{112}	11	S_{122}	12	S_{132}	13	S_{212}	14	S_{222}	15	S_{232}	16	S_{312}	17	S_{322}	18	S_{332}
C_3	19	S_{113}	20	S_{123}	21	S_{133}	22	S_{213}	23	S_{223}	24	S_{233}	25	S_{313}	26	S_{323}	27	S_{333}

Anmerkung: Die Bedingungen sind in blauer Schriftfarbe nummeriert. Die Faktoren A, B, C stellen die UVn dar, deren Stufen durch die jeweilige tiefergestellte Ziffer gekennzeichnet sind. S 111 bis S 333 sind die 27 experimentellen Bedingungen bzw. Teilstichproben. Beispiel: In Bedingung S 123 wird UV 1 als geringe Anzahl Downloads, UV 2 als hohe Anzahl Zitationen (Werk) variiert, UV 3 besitzt die Stufe „keine Angabe“.

4.1.5.1 Randomisierung der Surrogate und Aufgaben

Die Reihenfolge der zu bewertenden neun Surrogate in den drei Aufgaben wurde auch hier mithilfe des Online-Tools *Research Randomizer* ermittelt, indem ein Set aus 27 einmalig vorkommenden Zahlen von 1 bis 27 generiert wurde. Das Ergebnis der randomisierten Reihenfolge zeigt Tabelle 4.6; die manuelle Übertragung dieser Reihenfolge auf die einzelnen Bedingungen resultierte in den in Tabelle 4.7 enthaltenen Werten. Die Surrogate 1 bis 9 wurden in Aufgabe 1, die Surrogate 10 bis 18 in Aufgabe 2 und die restlichen Surrogate 19 bis 27 in Aufgabe 3 durch die Versuchspersonen bewertet. Die Festlegung, welches Surrogat welche Bedingung abdeckt, ist demzufolge für jede Versuchsperson dieselbe, d. h., das Surrogat, für das alle Stufen die geringe Anzahl (Stufe 1) aufweisen, ist immer Surrogat Nr. 16 in Aufgabe 2, jedoch nicht auf derselben Position.

Das Experiment umfasst die Bearbeitung der drei Aufgaben durch die Versuchspersonen in randomisierter Reihenfolge, wobei die Surrogate pro Aufgabe ebenfalls in zufälliger Reihenfolge angezeigt wurden. Zu beachten ist hierbei, dass die Surrogate nicht über die drei Aufgaben hinweg, sondern jeweils immer innerhalb derselben Aufgabe randomisiert wurden. Dieser Ablauf ist schematisch in Abbildung 4.2 dargestellt.

Die Reihenfolge der Anzeige der zu bewertenden Surrogate und der Aufgaben wurde den Merkmalen eines echten Experiments entsprechend während der Bearbeitung des Online-Fragebogens ad hoc durch das Fragebogentool (vgl. Abschnitt 4.2.3) randomisiert, damit die Ergebnisse keinen Reihen- bzw. Positionseffekten unterliegen. Solche Effekte sind bei der Art der Darstellung der zu bewertenden Surrogate zu berücksichtigen, denn Studien konnten zeigen, dass die Relevanzbewertung eines Dokuments durch andere Dokumente, die in dem Ergebnisset bereits zuvor gesehen und bewertet wurden, beeinflusst wird. Diese

Abbildung 4.1
Versuchsplan in der Notation der Campbell-Tradition

R	X_{A1B1C1}	O
R	X_{A1B2C1}	O
R	X_{A1B3C1}	O
R	X_{A2B1C1}	O
R	X_{A2B2C1}	O
R	X_{A2B3C1}	O
R	X_{A3B1C1}	O
R	X_{A3B2C1}	O
R	X_{A3B3C1}	O
R	X_{A1B1C2}	O
R	X_{A1B2C2}	O
R	X_{A1B3C2}	O
R	X_{A2B1C2}	O
R	X_{A2B2C2}	O
R	X_{A2B3C2}	O
R	X_{A3B1C2}	O
R	X_{A3B2C2}	O
R	X_{A3B3C2}	O
R	X_{A1B1C3}	O
R	X_{A1B2C3}	O
R	X_{A1B3C3}	O
R	X_{A2B1C3}	O
R	X_{A2B2C3}	O
R	X_{A2B3C3}	O
R	X_{A3B1C3}	O
R	X_{A3B2C3}	O
R	X_{A3B3C3}	O

Tabelle 4.6 Randomisierte Reihenfolge der Bedingungen

	Aufgabe 1		Aufgabe 2		Aufgabe 3
Surrogat 1	15	Surrogat 10	7	Surrogat 19	17
Surrogat 2	26	Surrogat 11	20	Surrogat 20	13
Surrogat 3	16	Surrogat 12	25	Surrogat 21	3
Surrogat 4	18	Surrogat 13	8	Surrogat 22	9
Surrogat 5	21	Surrogat 14	11	Surrogat 23	23
Surrogat 6	24	Surrogat 15	2	Surrogat 24	10
Surrogat 7	12	Surrogat 16	1	Surrogat 25	5
Surrogat 8	6	Surrogat 17	4	Surrogat 26	19
Surrogat 9	27	Surrogat 18	22	Surrogat 27	14

Tabelle 4.7 Übertragung der randomisierten Reihenfolge auf die Bedingungen

	Aufgabe 1		Aufgabe 2		Aufgabe 3
Surrogat 1	S232	Surrogat 10	S311	Surrogat 19	S322
Surrogat 2	S323	Surrogat 11	S123	Surrogat 20	S212
Surrogat 3	S312	Surrogat 12	S313	Surrogat 21	S131
Surrogat 4	S332	Surrogat 13	S321	Surrogat 22	S331
Surrogat 5	S133	Surrogat 14	S122	Surrogat 23	S223
Surrogat 6	S233	Surrogat 15	S121	Surrogat 24	S112
Surrogat 7	S132	Surrogat 16	S111	Surrogat 25	S221
Surrogat 8	S231	Surrogat 17	S211	Surrogat 26	S113
Surrogat 9	S333	Surrogat 18	S213	Surrogat 27	S222

Studien befassen sich mit den Reihenfolgeneffekten (*order effects)* und der Frage, wie die Reihenfolge der angezeigten Dokumente (z. B. Eisenberg & Barry, 1988; Purgailis Parker & Johnson, 1990; Scholer et al., 2013; Xu & Wang, 2008) und der zu bearbeitenden Aufgaben in IIR-Studien (Clemmensen & Borlund, 2016), aber auch inwieweit die Anzahl der präsentierten Dokumente (M. Huang & Wang, 2004) die Bewertung beeinflusst.[14]

[14] Saracevic (2016b, S. 67 ff.) fasst die Erkenntnisse aus diesen Untersuchungen unter der Überschrift „Beyond independence“ zusammen, was als bewusst gesetztes Signal gegen die Annahme im Cranfield-Paradigma gedeutet werden kann: „*The relevance of one document is independent of the relevance of any other document*“ (Buckley & Vorhees, 2005, S. 68).

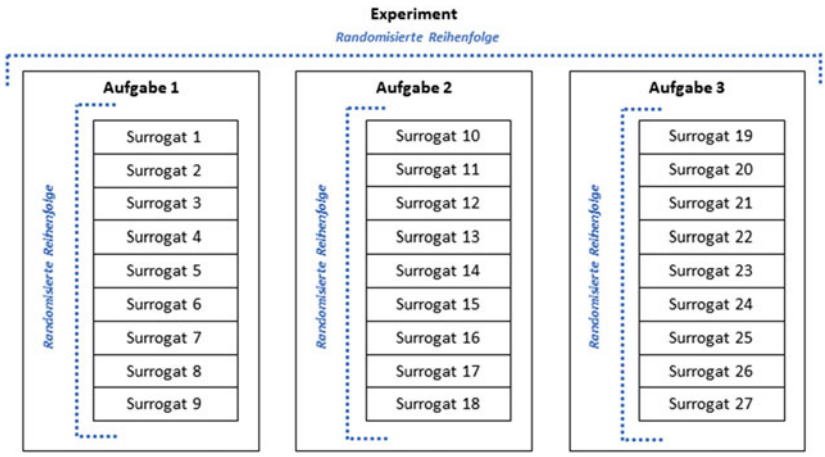

Abbildung 4.2 Ablauf des Experiments in schematischer Darstellung

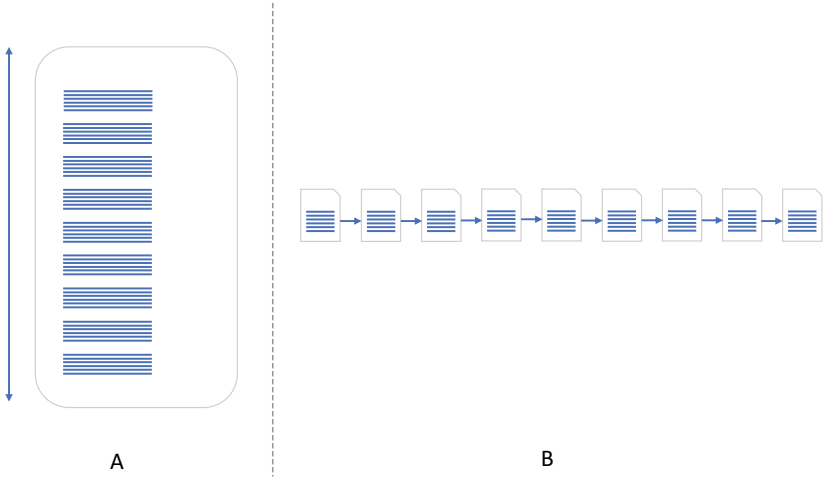

Abbildung 4.3 Möglichkeiten der Surrogate-Darstellung: (A) links in einer gemeinsamen Liste, (B) rechts separat in einer Reihe

In der traditionellen Information Retrieval-Evaluierung werden die zu bewertenden Dokumente den Jurorinnen und Juroren einzeln und nacheinander vorgelegt – eine Darstellung, die der Suchergebnispräsentation in heutigen (wissenschaftlichen) IR-Systemen widerspricht. Ob in dem hier beschriebenen Experiment die neun Surrogate jeder Aufgabe entweder (A) untereinander in einer Liste auf einer Seite oder (B) einzeln und nacheinander auf mehreren Seiten, also in einer Reihe, präsentiert werden (Abbildung 4.3), stellt somit eine wichtige Entscheidung dar. Darstellungsart (A) erlaubt ein Vor- und Zurückscrollen sowie die Bewertungen in einer flexiblen Reihenfolge vorzunehmen, (B) bietet diese Möglichkeit nicht, ohne zwischen einzelnen Seiten navigieren zu müssen, wodurch nie alle oder mehrere Surrogate zusammen betrachtet werden können.

Die Entscheidung für Darstellungsart (A), also eine Listendarstellung der Surrogate pro Aufgabe, und gegen eine separate Reihendarstellung wurde maßgeblich getroffen auf der Basis der Erkenntnisse über Effekte, die bei der Erhebung von Relevanzbewertungen auftreten können. Zur besseren Nachvollziehbarkeit über die Entscheidungsfindung werden diese Effekte im nachfolgenden Abschnitt vorgestellt und im Kontext des Experiments bewertet.

4.1.5.2 Effekte bei der Erhebung von Relevanzbewertungen

Der Prozess der Relevanzbewertung wird durch verschiedene system-, nutzer- und situationsbasierte Faktoren beeinflusst (vgl. Abschnitt 3.2.2). Ferner zeigen Studien, dass Relevanzbewertungen Effekten unterliegen, die bei der Erhebung von expliziten Relevanzbewertungen im Forschungskontext berücksichtigt werden sollten. Diese Effekte sind als ein Zusammenspiel aufzufassen, zum einen bedingt durch die Reihenfolge von (a) experimentellen Bedingungen, (b) zu bearbeitenden Aufgaben und (c) Dokumenten zur Relevanzbewertung, zum anderen durch individuelle Faktoren.

Der Prozess der Relevanzbewertung ist ein Beurteilungsprozess (vgl. Abschnitt 3.1.3), daher ist es sinnvoll, Effekte bei der Erhebung mit Blick auf Urteilsfehler, wie sie beispielsweise beim Einsatz von Ratingskalen auftreten können (Döring & Bortz, 2016, S. 252 ff.), zu betrachten. Einer dieser Effekte ist der Primacy-Recency-Effekt und bezeichnet als Sammelbegriff die Reihenfolgen- bzw. Positionseffekte, die bei der Beurteilung von sequenziell dargebotenen Objekten auftreten können: bei einer Bevorzugung der Objekte auf den Anfangspositionen handelt es sich um den Primacy-Effekt (Primäreffekt); werden Objekte auf den Endpositionen höher gewichtet, spricht man vom Recency-Effekt (Rezenzeffekt) (Döring & Bortz, 2016, S. 254–255). Bezogen auf die Bewertung von Dokumenten beschreiben Xu & Wang (2008) den Reihenfolgeneffekt dadurch, dass die Relevanz eines Dokuments unterschiedlich wahrgenommen

wird, wenn es an verschiedenen Positionen innerhalb einer Reihe präsentiert wird. In diesem Zusammenhang definieren sie die Begriffe *order effect*, *primacy effect* und *recency effect* wie folgt:

> [W]e define an order effect as a user's different evaluations towards a document when it is placed in different positions in a list. We define a primacy (recency) effect as the situation when a document is more favorably evaluated when it is placed earlier (later) in a list than when it is placed later (earlier). Primacy and recency effects are two types of order effect outcome. (Xu & Wang, 2008, S. 1266)

Abbildung 4.4 zeigt den Zusammenhang über die Effekte bei der Erfassung von expliziten Relevanzbewertungen basierend auf den Erkenntnissen von informationswissenschaftlichen Studien zu *order effects*: Reihenfolgeneffekte zeigen sich in Experimenten anhand von Ermüdung und Lerneffekten, die wiederum unerwünschte Effekte im Verhalten der Versuchspersonen nach sich ziehen können. Im Kontext der Relevanzbewertung untersuchten Studien Effekte, die ebenfalls auf Ermüdung und Lerneffekte zurückzuführen sind und mit Modellen über menschliches Verhalten erklärt werden können. Nachfolgend werden diese Effekte und Modelle kurz erläutert und die Maßnahmen, die diesen Erkenntnissen zufolge für die vorliegende Studie ergriffen wurden, dargelegt.

In Experimenten sind Ermüdung der Probandinnen und Probanden und mögliche Lerneffekte bei der Entwicklung des Untersuchungsdesigns zu berücksichtigen. Ermüdung (*fatigue*) kann die Ausführung von Tätigkeiten und die Leistung von Personen beeinträchtigen. Dies muss aber nicht zwangsläufig der Fall sein, der Effekt lässt sich beispielsweise durch Enthusiasmus oder in Notsituationen neutralisieren (Clemmensen & Borlund, 2016). Ermüdung wird beeinflusst durch die Anzahl, die Komplexität und den Schwierigkeitsgrad der zu bearbeitenden Aufgaben. Allerdings stellten Clemmensen & Borlund (2016) in ihrer Studie über den *order effect* in der IIR-Evaluierung fest: „[F]atigue is present, but the effect of it [on performance] is absent" (S. 210). Ein Lerneffekt kommt durch die kognitive Verarbeitung der wahrgenommenen Objekte zustande und, bezogen auf den Prozess der Informationssuche, durch die Vermehrung des Wissensstands innerhalb dieses Prozesses (siehe psychologische Relevanz, vgl. Abschnitt 3.1.1).

Aufgrund der Reihenfolge, in der Versuchspersonen die verschiedenen Bedingungen in einem Experiment mit Within-Subjects-Design durchlaufen, entstehen *Carry-over*-Effekte, die eine inhaltliche Beeinflussung der Teilnehmenden in einer Bedingung durch die vorangegangene Bedingung bezeichnen (Sedlmeier &

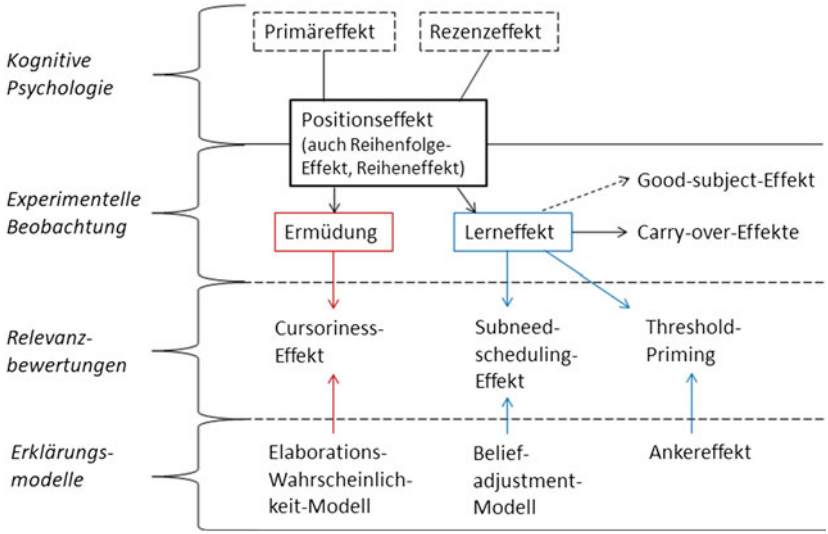

Abbildung 4.4 Effekte bei der Erhebung von expliziten Relevanzbewertungen

Renkewitz, 2018, S. 168). Solche Carry-over-Effekte treten in Between-Subjects-Designs nicht auf, da jede Versuchsperson ausschließlich einer Bedingung ausgesetzt ist.

Ein weiterer Effekt, der sich aufgrund von Gelerntem durch eine bestimmte Reihenfolge von Aufgaben bzw. Fragen zeigen kann, ist der *Good-subject*-Effekt (Clemmensen & Borlund, 2016). Die „gute Versuchsperson" ist geneigt sich erwartungskonform zu verhalten, also in einer Weise zu reagieren, die dazu führt, dass die Ergebnisse die vermutete Hypothese stützen.[15]

Ermüdung kann bei der Erhebung von Relevanzbewertungen den Effekt hervorrufen, dass Versuchspersonen Dokumente hinsichtlich ihres Inhalts auf Basis von vermeintlich eher unbedeutenden, nachrangigen Merkmalen (*peripheral cues*)[16] wie des Titels beurteilen; Xu & Wang (2008) bezeichnen diesen Effekt

[15] Der Good-subject-Effekt wird bei Clemmensen & Borlund (2016) als Folge eines Lerneffekts betrachtet, obwohl Versuchspersonen bereits zu Beginn einer Untersuchung gewisse Erwartungen mitbringen; das bedeutet, dass dieser Effekt nicht zwangsläufig und exklusiv als Folge eines Lerneffekts auftritt, daher ist er in Abbildung 4.4 mit einem gestrichelten Pfeil vom Lerneffekt ausgehend dargestellt.

[16] Popularitätsdaten können ebenfalls als derartige *peripheral cues* betrachtet werden, sollen sie doch als Indikator für die (inhaltliche) Qualität eines Werks dienen (vgl.

als *cursoriness effect*[17], untersuchten diesen in ihrer Studie jedoch nicht individuell, weshalb keine genaueren Aussagen über den Einfluss dieses Effekts getroffen werden können. Die Autoren begründen allerdings den Primacy-Recency-Effekt mit dieser Art der „flüchtigen" Bewertung:

> The cursoriness effect has two consequences: First, documents of lower average subjective relevance receive better evaluations and demonstrate a recency effect; second, documents of higher subjective relevance receive worse evaluations and demonstrate a primacy effect. (Xu & Wang, 2008, S. 1268).

Als Erklärungsmodell für den *cursoriness effect* verweisen Xu & Wang (2008) auf das Elaborations-Wahrscheinlichkeit-Modell (*Elaboration Likelihood Model*)[18]: So würden Motivation und kognitive Kapazitäten bestimmen, ob ein Informationsobjekt im Prozess der Informationsverarbeitung ausführlich und detailliert evaluiert wird oder zügig auf der Basis peripherer Hinweisreize. Mit Voranschreiten des Prozesses nehmen die kognitiven Kapazitäten der Versuchspersonen ab und Ermüdung tritt ein. Allerdings beziehen die Autoren sich auf die Bewertung von Dokumenten im Kontext der Websuche, wodurch Webseiten als Volltextdokumente dienen und eine Bewertung anhand des eigentlichen Inhalts (*evaluative judgments*) tatsächlich ermöglicht wird. Insofern stellt sich die Frage, inwieweit sich diese Theorie generell auf die Bewertung von Suchergebnissen in Form von Surrogaten (*predictive judgments*) und, mit Blick auf die vorliegende Arbeit, speziell auf Suchergebnisrepräsentationen in akademischen Suchsystemen übertragen lässt. Dieser Forschungsfrage kann im Rahmen dieser Arbeit nicht nachgegangen werden. Für das hier vorgestellte Experiment erschien die folgende Annahme plausibel: Der Cursoriness-Effekt tritt eher bei der Beurteilung von Informationsobjekten auf, die über den Umfang von Surrogaten in wissenschaftlichen Suchsystemen hinaus den eigentlichen Inhalt (zumindest teilweise) abbilden; auch das Abstract ist für eine detaillierte Evaluierung des Dokumenteninhalts nicht ausreichend.

Abschnitt 2.1.2.2). Ob es einen Zusammenhang zwischen Ermüdung und der Bewertung anhand solcher *peripheral cues* gibt, kann anhand der vorliegenden Studie nicht erklärt werden; eine derartige Fragestellung setzt den Faktor Ermüdung als zu variierende unabhängige Variable voraus.

[17] Dieser Effekt kann im Deutschen beispielhaft als Flüchtigkeitseffekt übersetzt werden.

[18] Das Elaborations-Wahrscheinlichkeit-Modell wurde von Petty & Cacioppo (1986) eingeführt als „general framework for organizing, categorizing, and understanding the basic processes underlying the effectiveness of persuasive communications" (S. 125).

Im Gegensatz zum ermüdungsbedingten Cursoriness-Effekt führt der Lerneffekt bei der Erhebung von Relevanzbewertungen zu zwei weiteren Effekten: dem Subneed-scheduling-Effekt und dem Threshold-Priming.

Den *subneed scheduling effect* beschreiben Xu & Wang (2008) als Anpassungseffekt, der durch die Bewertung anhand eines *subneed*, d. h. eines Teilbedürfnisses des Informationsbedürfnisses einer informationssuchenden Person, auftritt in Abhängigkeit von dem Zeitpunkt des Erscheinens eines zu bewertenden Dokuments und somit auch in Abhängigkeit seiner Position: Erscheint ein potenziell relevantes Dokument zu früh, wird es als thematisch weniger relevant erachtet; erscheint es zu spät, wird es ebenfalls als thematisch weniger relevant aber auch mit einem geringeren Neuigkeitswert beurteilt. Im erstgenannten Fall kann die thematische Relevanz nicht „korrekt" beurteilt werden, weil die informationssuchende Person ihren Wissensstand auf der Basis der ersten Dokumente erweitert und ihr Informationsbedürfnis anpasst oder konkretisiert; im letztgenannten Fall passen die Dokumente in ihrer thematischen Relevanz nicht mehr aus demselben Grund und sie haben ihren Neuigkeitswert verloren. Dieser Anpassungseffekt konnte empirisch nachgewiesen werden bei einem Listenumfang von circa 40 Dokumenten (Xu & Wang, 2008).

Wie im Zusammenhang mit dem Cursoriness-Effekt angeführt, wurden in der Studie Webseiten bewertet, wodurch eine Bewertung des Suchergebnisses über die Metadaten (das Surrogat) hinaus erfolgen konnte. Vor diesem Hintergrund ist auch das von den Autoren als Erklärungsmodell für den Subneed-scheduling-Effekt angebotene Belief-adjustment-Modell[19] kritisch zu betrachten. Es basiert auf der Annahme, dass ein gegenwärtiger Wahrnehmungszustand mit dem Hinzukommen neuer Informationen angepasst wird und weist somit eine große Ähnlichkeit mit dem ASK-Konzept auf (Clemmensen & Borlund, 2016). Gemeint ist, dass spätere Informationen weniger Aufmerksamkeit erhalten und Anpassungen an diese zu einem verzerrten Einfluss früherer Informationen auf den finalen Zustand führen:

> This type of order effect is known as the primacy effect in information integration. Similarly, if a situation leads to more cognitive attention to the later items, a recency effect emerges, with later items being more significant in the final belief. (Xu & Wang, 2008, S. 1266)

[19] Das Modell wurde von Hogarth & Einhorn (1992) entwickelt mit der Annahme: „Our theory assumes that people handle belief-updating tasks by a general, sequential anchoring-and-adjustment process in which current opinion, or the anchor, is adjusted by the impact of succeeding pieces of evidence" (S. 8).

Die Frage, die sich daraus ergibt, lautet, wie umfangreich oder inhaltlich detailliert diese neu hinzukommenden Informationen beschaffen sein müssen, damit sich der Wissensstand der informationssuchenden Person in dem Maße vergrößern kann, sodass das *subneed* tatsächlich angepasst wird. Auch die Bearbeitung dieser Frage würde den Rahmen dieser Arbeit sprengen.

Das Threshold-Priming beschreiben Scholer u. a. (2013) als Effekt, der durch unterschiedliche Kalibrierungen interner Relevanzmodelle (Heuristiken) bei menschlichen Jurorinnen und Juroren auftritt. So würden heuristische Relevanzmodelle gebildet und Relevanzstufen (*relevance thresholds*) im Kontext der Bewertung beeinflusst durch die Bewertung (des Relevanzgrades) des zuvor gesehenen Dokuments. – „A long sequence of irrelevant documents, for instance, might cause an assessor to lower their threshold of relevance, or alternatively to lose concentration and miss relevant documents“ (Scholer et al., 2013, S. 623). Dieser Effekt kann demnach ebenfalls als Anpassungseffekt betrachtet werden. Ähnlich dem Belief-adjustment-Modell weisen die Autoren auf den Ankereffekt[20] als Erklärungsmodell hin. Hierbei handelt es sich um eine kognitive Verzerrung, welche die menschliche Neigung erklärt, sich bei der Urteils- und Entscheidungsfindung auf zuerst präsentierte Informationen (Anker) zu beziehen. Die Ergebnisse der Studie von Scholer u. a. (2013) zeigen, dass in dem Fall, dass Juroren zu Beginn mit hochrelevanten Dokumenten konfrontiert werden, sie dazu tendieren, späteren Dokumenten niedrigere Relevanzbewertungen zu geben (und umgekehrt). In ihrer Studie zum Ankereffet bei Relevanzbewertungen konnten Shokouhi u. a. (2015) allerdings ein anderes Bewertungsverhalten beobachten: Die Juroren tendierten dazu, einem Dokument die gleiche Relevanzbewertung wie dem Dokument zu geben, das sie unmittelbar davor gesehen hatten. Die Autoren erklären diesen Unterschied mit den verschiedenen Arten von Ankereffekten. Während Scholer u. a. (2013) einen langfristigen Ankereffekt untersuchten, betrachteten Shokouhi u. a. (2015) einen kurzfristigen Ankereffekt:

> The differences are caused by the type of anchoring that is considered in the two studies. Scholer et al. focus on long-term anchoring (top k labeled documents as the anchor) and analyze how this affects the relevance labels assigned to the documents judged later. In our work, we focus on the short-term anchoring (last labeled document as the anchor) and analyze how this affects the relevance labels assigned to the document judged immediately after. (Shokouhi et al., 2015, S. 964)

[20] Tversky & Kahneman (1974) erforschten Heuristiken und kognitive Verzerrungen in der Beurteilung von Informationen und stellten unter anderem fest, dass der Ankeraffekt nicht nur bei Laien auftritt, sondern auch bei Experten als Folge von intuitivem Denken.

Scholer u. a. (2013) ließen die Dokumente im Stil von TREC einzeln und nacheinander (seitenweise) von Juroren bewerten; Shokouhi u. a. (2015) beschreiben die Form der Dokumentenpräsentation für die Juroren in ihrer Studie zwar nicht ausdrücklich, es ist aber davon auszugehen, dass diese die Dokumente ebenfalls einzeln und nacheinander sahen, da nur so sichergestellt werden konnte, dass ein Dokument zunächst wahrgenommen, dann bewertet und im Anschluss das nachfolgende Dokument betrachtet wird. Mit einer Präsentation der Dokumente als gemeinsame Liste hätte der Einfluss des Ankereffekts nicht untersucht werden können.

Zusammengefasst sind die Effekte, die bei der Erhebung von Relevanzbewertungen in einem Forschungskontext auftreten können, insbesondere auf Ermüdung und Lerneffekte zurückzuführen. In Experimenten wird versucht, Lerneffekte (Störvariablen) zu kontrollieren, z. B. durch Randomisieren, Ausbalancieren und die Durchführung als Blindversuch. Maßnahmen im Umgang mit potenziellen Störvariablen im Rahmen des hier berichteten Online-Experiments wurden bereits in Abschnitt 4.1.4 erläutert. Um einer Ermüdung der Versuchspersonen entgegenzuwirken, sollten Anzahl und Schwierigkeitsgrad der zu bearbeitenden Aufgaben genau abgewägt werden und neben dem Zeitaufwand auch die Motivation der Versuchsperson in den Blick genommen werden. So sind Anreize zu schaffen, die potenzielle Versuchspersonen zunächst zur Teilnahme anregen (vgl. Abschnitt 4.2.5) und schließlich die Teilnehmenden zur Weiterbearbeitung und zum Beenden der Untersuchung motivieren. Demnach sind extrinsische wie intrinsische Faktoren bei der Motivation von Bedeutung.

Motivation kann durch Anregung von Interesse oder einen Neuigkeitswert hervorgerufen werden. Diesbezüglich empfehlen Clemmensen & Borlund (2016), die zu bearbeitenden Aufgaben entsprechend zu gestalten und beispielsweise das Konzept der *Simulated work task situations* umzusetzen. Ihrer Empfehlung wurde in dem hier vorgestellten Experiment gefolgt, da die Beschreibungen der Informationsbedürfnisse auf dem genannten Konzept basieren (vgl. Abschnitt 4.2.1). Einen intrinsischen Motivationsfaktor stellt das *Need for Cognition*[21] (NfC) dar. Es kennzeichnet die Bereitschaft von Menschen, ein hohes Maß an kognitiven Kapazitäten bei der Bearbeitung von Aufgaben bzw. der Lösung von Problemen aufzuwenden und dabei Vergnügen zu empfinden; kurz bedeutet es die Freude

[21] Das Konzept des Need for Cognition (NfC) definieren Cacioppo & Petty (1982) als „[P]eople's tendency to engage in and enjoy thinking" (S. 130). Das Messinstrument zur Erfassung des NfC wurde ursprünglich mit 34 Items entwickelt; inzwischen ist eine kürzere Fassung mit 18 Items üblich, die Aussagen zur Beurteilung enthält wie „I prefer complex to simple tasks" oder „Thinking is not my idea of fun" als Beispiel eines negativ gepolten Items (Petty, Briñol, Loersch, & McCaslin, 2014, S. 319).

am Denken. Neben dem Einfluss des Ankereffekts untersuchten Scholer u. a. (2013) auch den Einfluss des NfC. In ihrer Studie stellten sie fest, dass Personen mit einem niedrigeren NfC niedrigere Relevanzbewertungen vergaben und Personen mit einem höherem NfC mehr Zeit auf Relevanzbewertungen verwendeten und eher mit den Expertenbewertungen überein stimmten[22]. Da die Zielgruppe der vorliegenden Studie in der Wissenschaft tätige Akademikerinnen und Akademiker umfasst, ist bei den Teilnehmenden von einem hohen NfC auszugehen. Motivationsfaktoren sind dabei eher als Einflussparameter in Hinblick auf das erfolgreiche Abschließen des Fragebogens (das Durchhalten der Teilnehmenden) zu sehen, da personenbezogene Störvariablen mit der Durchführung des Experiments als Within-Subjects-Design ohnehin perfekt ausbalanciert sind und sich derartige Effekte in den Ergebnissen nicht wiederfinden sollten.

Abschließend lässt sich die Frage, ob im Rahmen des hier vorgestellten Experiments die Surrogate einer Aufgabe als gemeinsame Liste untereinander oder einzeln und nacheinander den Versuchspersonen präsentiert werden sollten, mit Blick auf den Subneed-Scheduling-Effekt und das Threshold-Priming beantworten: Beide Effekte wurden in Studien untersucht, in denen die Juroren keine Listendarstellung der zu bewertenden Dokumente erhielten. Für das Experiment im Rahmen dieser Studie wurde eine Listendarstellung gewählt, um solche Effekte zu verhindern. Bei einer gemeinsamen Darstellung der Surrogate in einer Liste werden diese nicht separat wahrgenommen, sondern können im Kontext der anderen Ergebnisse betrachtet und bewertet werden. Aufgrund der Möglichkeit, ein nachträgliches Anpassen der Bewertung mit dem Schieberegler oder das Bewerten in einer von den Versuchspersonen frei gewählten Reihenfolge vorzunehmen, ließen sich eventuelle Auffälligkeiten in den Ergebnisdaten nicht eindeutig auf die genannten Effekte zurückführen.

Durch die zusätzliche Randomisierung der Reihenfolge der angezeigten Surrogate lassen sich mögliche Reihenfolge- bzw. Positionseffekte aus den Ergebnissen herausrechnen. Dies ist zwar unabhängig von einer Entscheidung für oder gegen

[22] In ihrer Studie teilten Scholer u. a. (2013) die Bewertungen der Juroren in zwei Gruppen – eine Gruppe enthielt die Bewertungen der Juroren mit niedrigem NfC-Wert, die andere Gruppe die Bewertungen der Juroren mit einem hohen NfC-Wert. Die Aufteilung erfolgte anhand des Medians und ist kritisch zu betrachten, da dieser eine eher künstliche Grenze darstellt und unklar bleibt, ab welchem Wert ein NfC als hoch eingestuft werden darf. Eine Aufteilung der Bewertungen in drei Gruppen und ein Vergleich der beiden Randgruppen ohne Berücksichtigung der Werte in der mittleren Gruppe erscheint hierbei geeigneter. Aufgrund der fehlenden Erfassung des NfC im Rahmen der hier beschriebenen Studie und der fehlenden Expertenbewertungen als Goldstandard im Stil von TREC kann ein derartiger Vergleich nicht vorgenommen werden; zudem ist für die Beantwortung der Forschungsfragen eine solche Untersuchung irrelevant.

eine Listendarstellung, sei um die Vollständigkeit der Argumente willen dennoch an dieser Stelle erwähnt.

Schließlich birgt die Listendarstellung den Vorteil, den realen Suchergebnisrepräsentationen in modernen akademischen Suchsystemen zu entsprechen und den Teilnehmenden eine dahingehend weniger künstliche oder ungewohnte Suchumgebung bieten zu können.

4.2 Planung und Umsetzung der Datenerhebung

Wie bei vielen bisherigen Studien zur Erforschung von Relevanzkriterien (vgl. Abschnitt 2.2) erfolgte die Datenerhebung durch Befragung der Versuchspersonen, wobei die Befragung bei dieser Studie die Erhebung von expliziten Relevanzbewertungen in den Fokus nimmt.

Konkret handelt es sich bei der Befragungsmethode um einen multifaktoriellen Online-Survey, der auch als (Online-)Vignettenanalyse, bezeichnet wird. Vignetten sind kurze Beschreibungen von Situationen mit variierenden Merkmalen, die von den Teilnehmenden beurteilt werden (Berger & Wolbring, 2015, S. 46). In dem vorliegenden Experiment können die Beschreibungen der Informationsbedürfnisse als Vignetten gesehen werden, wobei die Variationen in den zu bewertenden Surrogaten mit den manipulierten Popularitätsdaten liegen. Multifaktorielle Surveys stellen allerdings das einzige Experiment dar, in dem kein tatsächliches Verhalten, sondern Verhaltensintentionen wie Einstellungen oder Präferenzen erfasst werden, ohne dass Konsequenzen zu erwarten wären (Berger & Wolbring, 2015, S. 46). Das Bewerten von Surrogaten bezüglich ihrer Relevanz (Nützlichkeit) zur Befriedigung des Informationsbedürfnisses im Rahmen des Experiments kann als eine Erfassung von Verhaltensintentionen (Auspurg & Hinz, 2007, S. 295) gesehen werden. Diese Auffassung geht mit der Annahme einher, den Prozess der Relevanzbewertung von Suchergebnissen als einen Beurteilungsprozess zu begreifen und stellt ein inhaltliches Argument für die Wahl eines experimentellen Designs in Form eines (multi)faktoriellen Surveys zur Erforschung von Kriterien bei der Relevanzbewertung von Surrogaten dar.

In den nachfolgenden Abschnitten wird dargelegt, wie die Umsetzung des experimentellen Designs (vgl. Abschnitt 4.1) für die Datenerhebung konkret erfolgte. Die dem entwickelten Modell zur subjektiven Relevanzbewertung von Suchergebnissen in akademischen Suchsystemen zugrunde gelegte Relevanzdefinition (vgl. Abschnitt 3.1.1) schlägt sich nicht nur in der Operationalisierung der abhängigen Variable nieder, sie steuert ebenfalls die Darstellungsart für die Datenerhebung: So wird der ziel- und aufgabenorientierten Relevanzdefinition

dadurch Rechnung getragen, dass im Online-Fragebogen die Beschreibungen der Informationsbedürfnisse den zu bewertenden Surrogaten vorangestellt sind, um die Versuchspersonen als erstes auf die Kontextinformationen hinzuweisen, bevor sie durch Herunterscrollen die Surrogate sehen[23]. Ferner beschränkt sich die Bewertung auf Surrogate, also auf *predictive judgments*, die die Einbindung oder Zugänglichmachung von Volltexten durch Verlinken verbietet. Diese und weitere Entscheidungen, die bezüglich der Herkunft der als Vorlage dienenden Dokumente für die Erstellung der Surrogate sowie der Art der Darstellung innerhalb der Suchergebnisliste getroffen wurden, beschreibt Abschnitt 4.2.2. Zunächst wird in Abschnitt 4.2.1 die Entwicklung der Aufgaben, die jeweils mit der Beschreibung eines Informationsbedürfnisses beginnen, erläutert; das Konzept der *Simulated Work Task Situations* diente dabei als Vorlage (Abschnitt 4.2.1.1). Des Weiteren werden das Vorgehen bei der Auswahl der Themen für die Aufgaben beschrieben (Abschnitt 4.2.1.2) sowie die Beschreibungstexte der drei Aufgaben angeführt (Abschnitt 4.2.1.3).

Neben der Entwicklung der Beschreibungen der Informationsbedürfnisse und der Erstellung der Surrogate gingen mit der Konstruktion des Fragebogens (Abschnitt 4.2.3) weitere Überlegungen in Hinblick auf zusätzliche Fragen einher. So wurden zunächst demografische Angaben der Teilnehmenden erfasst (Abschnitt 4.2.3.1), bevor ein Vorabfragebogen Items zum Informationssuchverhalten zeigte (Abschnitt 4.2.3.2). Zusätzlich zum eigentlichen Experiment, also der Erfassung der expliziten Relevanzbewertungen (Abschnitt 4.2.3.3), wurden den Teilnehmenden auch im Anschluss weitere Fragen gestellt, deren Antworten unter anderem den Nachteil quantitativer Methoden mildern sollten, indem teilweise offene Fragen Freitexteingaben verlangten, die für eine Auswertung vorab zu codieren sind (Abschnitt 4.2.3.4). Der Fragebogen schloss mit der Aufklärung über den tatsächlichen Zweck der Studie und der Möglichkeit zur Teilnahme an einem Gewinnspiel (Abschnitt 4.2.3.5).

Die Zielgruppe dieses Online-Experiments sind Personen, die Kenntnisse über den wissenschaftlichen Publikationsprozess und praktische Erfahrung im Umgang mit akademischen Suchsystemen besitzen. Es ist davon auszugehen, dass dies auf Personen zutrifft, die in einem Umfeld tätig sind, in dem sie zu eigenen Forschungszwecken regelmäßig mit akademischen Suchsystemen interagieren. Als potenzielle Teilnehmende kommen daher Personen infrage, die bereits erfolgreich eine umfangreiche, wissenschaftliche Arbeit eigenständig verfasst haben,

[23] Je nach Bildschirmgröße bzw. Größe des Browserfensters kann zunächst nur der Beschreibungstext oder der Beschreibungstext und das erste von neun Surrogaten vollständig sichtbar und lesbar sein.

d. h. mindestens einen Studienabschluss als Master, Diplom[24] oder Magister besitzen und aktuell an Hochschulen oder außeruniversitären Forschungseinrichtungen affiliiert sind. Dies trifft zu auf Promovierende, wissenschaftliche Mitarbeiterinnen und Mitarbeiter, Professorinnen und Professoren.

Die Zugehörigkeit zu einer bestimmten wissenschaftlichen Fachdisziplin stellt in Bezug auf die unterschiedlichen Forschungskulturen und deren Zitationsverhalten einen weiteren Einflussfaktor[25] auf den Prozess der Relevanzbewertung von Surrogaten in wissenschaftlichen Suchsystemen dar (vgl. Abschnitt 3.2.2). Vor diesem Hintergrund verlangte die Entscheidung, ob die Auswahl der Teilnehmenden auf eine zuvor festgelegte Wissenschaftsdisziplin begrenzt oder unabhängig einer Disziplinzugehörigkeit sein sollte, eine sorgfältige Abwägung.

Letztlich wurde darauf verzichtet, die Stichprobe auf Personen einer bestimmten Fachdisziplin einzugrenzen, sondern es sollten Teilnehmende aus möglichst allen wissenschaftlichen Disziplinen gewonnen werden, um den optimalen Stichprobenumfang zu erzielen und erste Erkenntnisse über den generellen Einfluss von Popularitätsdaten als Bestandteil von Suchergebnissen auf die Relevanzbewertung zu erlangen. Die einzige Ausnahme bestand darin, dass die Probanden keinen bibliotheks- oder informationswissenschaftlichen Hintergrund mitbringen sollten. Begründet wird diese Ausnahme im Zusammenhang mit der Frage, welche Themen die Informationsbedürfnisse und Surrogate abdecken sollten. Ferner kommt der externen Validität des Experiments zugute, dass die Versuchspersonen nicht alle derselben Fachdisziplin oder derselben Statusgruppe (z. B. Studierende) angehören (Berger & Wolbring, 2015, S. 46). Die Stichprobe in dieser Studie gehört somit zur Population der Wissenschaftlerinnen und Wissenschaftler an Universitäten in Deutschland. Die Berechnung des erforderlichen Stichprobenumfangs erfolgt in Abschnitt 4.2.4, das Vorgehen bei der Gewinnung der Studienteilnehmenden wird in Abschnitt 4.2.5 ausführlich erläutert.

[24] Im Gegensatz zum Diplom (FH), das nach der Bologna-Reform mit dem Bachelor-Abschluss gleichgesetzt wird, ist hier das an Universitäten erworbene Diplom vor der Bologna-Reform gemeint.

[25] Die Fachdisziplin wurde in diesem Experiment nicht als eigene UV untersucht; die Zugehörigkeit zu einem Fach kann allerdings als personengebundene Störvariable betrachtet werden, die aufgrund des gewählten Within-Subjects-Designs kontrolliert ist (vgl. Abschnitt 4.1.4). Die Ergebnisse lassen sich jedoch auch nach Fachdisziplin getrennt betrachten, um Hinweise auf mögliche Gruppenunterscheide zu erhalten, die in zukünftigen Forschungsvorhaben gezielt untersucht werden könnten.

Vor der Durchführung der Studie wurde geprüft, ob eine formale Genehmigung von der Ethikkommission der Universität Hildesheim[26] einzuholen ist. Da die Probandinnen und Probanden während der Teilnahme keinen gesundheitlichen oder zu Beeinträchtigungen führenden Risiken ausgesetzt waren (Döring & Bortz, 2016, S. 130 f.), wurde darauf verzichtet, eine entsprechende Anfrage zu stellen. Ferner wurden forschungsethische sowie datenschutzrechtliche Bestimmungen im Zuge der Datenerhebung und auch bei der späteren Datenanalyse befolgt. Diese wurden bei der Fragebogenkonstruktion (vgl. Abschnitt 4.2.3) und der Probandenakquise (vgl. Abschnitt 4.2.5) unter anderem in Hinblick auf Freiwilligkeit und mit dem Hinweis auf die Speicherung personenbezogener Daten explizit berücksichtigt.

4.2.1 Entwicklung der Informationsbedürfnisse

Obwohl die Studie vom Untersuchungsdesign her dem klassischen Ansatz der Information Retrieval-Evaluierung folgt, wurde aus verschiedenen Gründen auf eine Nachnutzung bestehender Topics aus dem TREC- oder IIR-Kontext verzichtet. Eine Nachnutzung von TREC-Topics wurde hauptsächlich aus dem folgenden Grund ausgeschlossen: Wie bereits in mehreren Abschnitten erwähnt, wird die dem Cranfield-Paradigma folgende, traditionelle IR-Evaluierung dem dynamischen und kontextabhängigen Relevanzkonzept aus der subjektiven Perspektive einer informationssuchenden Person nicht gerecht. Dieses wird insbesondere anhand der inzwischen veralteten TREC-Aufgaben deutlich, denn diese enthielten neben allgemeinen Informationen zum Thema unter anderem eine explizite Definition eines als relevant zu bewertenden Dokuments, wie zum Beispiel die folgende Beschreibung aus den ersten Tracks TREC-1 und TREC-2 zeigt (Harman, 2005, S. 30):

[26] https://www.uni-hildesheim.de/organe-und-gremien/senat/kommissionen/ethikkommission/ (letzter Zugriff: 15.05.2020); Hinweise zur Antragstellung https://www.uni-hildesheim.de/media/forschung/Forschung/11_Ethik_und_Transparenz/Ethikkommission/Uni_Hildesheim_-_Ethikkommission_-_Hinweise_Antragstellung.pdf (letzter Zugriff: 15.05.2020).

<num> Number: 053
<dom> Domain: International Economics
<title> Topic: Leveraged Buyouts

<desc> Description:
Document mentions a leveraged buyout valued at or above 200 million dollars.

<smry> Summary:
Document mentions a leveraged buyout valued at or above 200 million dollars.

<narr> Narrative:
A relevant document will cite a leveraged buyout (LBO) valued at or above 200 million dollars. The LBO may be at any stage, e.g., considered, proposed, pending, a fact. The company (being) taken private must be identified. The offer may be expressed in dollars a share.

<con> Concept(s):
1. leveraged buyout, LBO
2. take private, go private
3. management-led leveraged buyout

<fac> Factor(s):
<price> Price: >= 200 million dollars </fac>

<def> Definition(s):
Leveraged Buyout (LBO) – Takeover of a company using borrowed funds, with the target company's assets serving as security for the loans taken out by the acquiring firm, which repays the loans out of the cash flow of the acquired company or from the sale of the assets of the acquired firm.

Das Aussehen der traditionellen TREC-Topics hat sich inzwischen verändert; so werden seit 2015 neue, elaborierte Topics eingesetzt, wie beispielsweise dieser Task zeigt (Quelle: https://trec.nist.gov/data/tasks/subtasks.txt):

Task id: 7

disneyland paris [I'm planning my visit to Disneyland Paris]

* Information about Disneyland Paris
* Disneyland Paris entrance fee
* Book a hotel
* Choose the right tickets and buy them
* Book flights/trains
* Avoid queues
* Plan your visit, what to do, when
* Plan meals and drinks in and out of the park

Diese neueren Tasks beinhalten zwar eine Ziel- bzw. Aufgabenbeschreibung, wie sie auch für den Einsatz in Interactive Information Retrieval (IIR)-Studien gefordert wird (Wildemuth et al., 2014), zeigen jedoch eher eine Auflistung von untergeordneten Aufgaben zu einem Thema, als dass sie Kontextinformationen über die Situation oder Motivation hinter dem eigentlichen Informationsbedürfnis böten. Für das hier beschriebene Online-Experiment wurde daher auf das Konzept der *Simulated Work Task Situations* zurückgegriffen, das einen narrativen Ansatz verfolgt. Tasks, die diesem Konzept folgen und bereits in IIR-Studien entwickelt und erprobt wurden, kamen für eine Nachnutzung in dem hier beschriebenen Experiment allerdings ebenfalls nicht infrage, da nicht alle Eigenschaften bestehender Tasks den Anforderungen hinsichtlich der Zielgruppe, des Suchkontexts und des Kontexts in Hinblick auf akademische Suchsysteme entsprachen.[27]

4.2.1.1 Simulated Work Task Situations

Das Konzept der simulierten Arbeitsaufgabensituation (*simulated work task situation,* SWTS) wurde von Borlund & Ingwersen (1997) eingeführt und später als wesentlicher Bestandteil in das von Borlund (2003b) entwickelte Framework zur Evaluierung von IIR-Systemen aufgenommen.

Dabei handelt es sich um kurze Beschreibungstexte von Informationsbedürfnissen im Kontext einer bestimmten Situation, die die Testpersonen zur Suche

[27] Eine systematische Auflistung von in IIR-Studien verwendeten Tasks bietet das RepAST – Repository of Assigned Search Tasks, das unter der URL https://ils.unc.edu/searchtasks/search.php frei zugänglich ist und in Zusammenarbeit mehrerer Universitäten entstand (Freund & Wildemuth, 2014).

in einem IR-System motivieren sollen (Borlund, 2003a). Zum einen dienen die Beschreibungstexte dazu, bei den Testpersonen ein simuliertes Informationsbedürfnis hervorzurufen, zum anderen liefern sie den Rahmen zur Bewertung von situativer Relevanz; konkret erhalten die Testpersonen Informationen über das Informationsbedürfnis, den Kontext der jeweiligen Situation, das hinter dem Bedürfnis stehende Informationsproblem sowie das zu erreichende Ziel der Suche (Borlund & Ingwersen, 1997).

Eine *simulated work task situation* sieht beispielsweise wie folgt aus (Borlund, 2016, S. 396):

> Simulated work task situation: after your graduation you will be looking for a job in industry. You want information to help you focus your future job seeking. You know it pays to know the market. You would like to find some information about employment patterns in industry and what kind of qualifications employers will be looking for from future employees.

Borlund (2016) definiert die folgenden fünf Voraussetzungen für den Einsatz von SWTSs (S. 406–407):

(1) To tailor the simulated work task situation to the test participants:
   - a situation the test participants can relate to and identify themselves with;
   - a situation the test participants find topically interesting and/or of relevance to them; and
   - a situation that provides enough imaginative context in order for the test participants to be able to apply the situation.

(2) To include test participants' personal information needs as baseline.
(3) To rotate the order of simulated work task situation and personal information needs (counterbalancing).
(4) To pilot test prior to actual testing (often more than once).
(5) To display the used simulated work task situations when reporting the study.

Diesen Voraussetzungen zufolge sollen Beschreibungen von SWTSs auf die Zielgruppe zugeschnitten sein, sodass die Situation für die Testpersonen authentisch, realistisch und relevant ist, was eine gewisse Homogenität der Zielgruppe voraussetzt (Borlund, 2016, S. 396 ff.). Sind die unter Punkt (1) genannten Voraussetzungen an die Entwicklungen der Aufgaben erfüllt, resultiert dies in einer sinnvollen Balance zwischen weitgehend realistischen Bedingungen einerseits und dem in Experimenten notwendigen Maß an Kontrolle andererseits, unter der Voraussetzung, dass die Aufgaben für alle Testpersonen (auch im Wortlaut) gleich sind, und wie in Punkt (3) gefordert, in unterschiedlicher Reihenfolge angezeigt werden (Borlund, 2016, S. 396).

Vor diesem Hintergrund stellt jedoch die Integration von persönlichen Informationsbedürfnissen der Testpersonen (Punkt 2) die experimentelle Kontrolle vor ein Problem: Dadurch, dass die Testpersonen jeweils zusätzlich ein eigenes Informationsbedürfnis mitbringen, können diese Daten als Baseline nur für die jeweilige Person individuell betrachtet werden; ein Vergleich mit den Daten der anderen Testpersonen bezüglich deren persönlicher Informationsbedürfnisse kann aufgrund der zu erwartenden Unterschiede hinsichtlich Komplexität und Schwierigkeitsgrad nicht getroffen werden.

Die Forderung nach einem Pretest oder mehreren Testläufen (Punkt 4) zielt auf die mögliche Anpassung insbesondere der Aufgabenbeschreibungen, welche durch die Erhebung qualitativer Daten (z. B. Interview, Methode des lauten Denkens, halb-standardisierter Fragebogen mit offenen Fragen) erfüllt werden kann. Punkt (5) bezieht sich auf die Verwertung der Forschungsergebnisse und verlangt, die den Testpersonen vorgelegten Aufgabenbeschreibungen als Bestandteil des Studiendesigns exakt abzubilden; auf diese Weise können sich Leserinnen und Leser der Publikation ein eigenes Urteil über die Angemessenheit der Aufgaben und die Validität der Ergebnisse bilden. Auch in Hinblick auf die Durchführung von Replikationsstudien ist diese Forderung sinnvoll und nützlich.

Das Konzept der SWTSs kann als ein inzwischen etabliertes Instrument bei der Datenerhebung in IIR-Studien gesehen werden, denn bereits in mehr als sechzig empirischen Studien wurden SWTSs verwendet, auch wenn nicht alle Anforderungen an deren Erstellung und Umsetzung im Untersuchungsdesign erfüllt wurden (Borlund, 2016).

Die vorliegende Forschungsarbeit befasst sich mit dem Relevanzkonzept, das im Kontext des Interactive Information Retrieval ein zentrales Thema darstellt (vgl. Kapitel 1), allerdings ohne, dass die Probanden im Rahmen des Experiments selbständig eine Suchanfrage formulierten und mit dem System interagierten. Obwohl die zu bearbeitenden Aufgaben ausschließlich in der Bewertung von Suchergebnissen bestanden, wurden für die Nutzerstudie Aufgaben entwickelt, die die Voraussetzungen für den Einsatz von *simulated work task situations* bestmöglich erfüllen, wie in Tabelle 4.8 aufgelistet. Lediglich auf die Einbindung persönlicher Informationsbedürfnisse der Probanden musste aufgrund des experimentellen Designs in Hinblick auf Manipulation und Kontrolle verzichtet werden.

Tabelle 4.8 Überprüfung der Voraussetzungen von SWTSs für das Online-Experiment

Voraussetzung		Erfüllung/Anpassung
1–1	Testpersonen können sich mit der Situation identifizieren	Angestrebt
1–2	Testpersonen finden das Thema interessant	Angestrebt
1–3	Testpersonen erhalten ausreichend Kontextinformationen	Angestrebt
2	Verwendung von persönlichen Informationsbedürfnissen der Testpersonen als Baseline	Nein*
3	Anzeige der Aufgaben in randomisierter Reihenfolge	Ja
4	Überprüfung durch Pretest(s)	Ja
5	Einfügen in Methodenbeschreibung bei Veröffentlichung der Studie	Ja

*Eine Ad-Hoc-Integration von weiteren Aufgaben ist nicht möglich, da das gewählte Untersuchungsdesign die Vorgabe der Aufgaben und zu bewertenden Suchergebnisse verlangt.

4.2.1.2 Themenauswahl

Die Auswahl der Themen war maßgeblich beeinflusst von der Population, die die Stichprobe repräsentiert. Die Teilnehmenden konnten unabhängig ihrer Disziplinzugehörigkeit mitwirken, als einzige Ausnahme waren Personen mit einem bibliotheks- oder informationswissenschaftlichen Hintergrund von der Teilnahme ausgeschlossen. Der Grund dafür bestand darin, dass die Informationsbedürfnisse Themen aus diesem Fachgebiet behandeln sollten. Das Ziel bestand darin, solche informationswissenschaftlichen Themen zu identifizieren, die auch für Personen mit einem nicht-informationswissenschaftlichen Hintergrund interessant und verständlich sind.

Die Motivation hinter dieser Entscheidung soll anhand eines Beispiels veranschaulicht werden: Bei einem biologischen Thema, wie beispielsweise dem Jagdverhalten von Katzen als möglicher Forschungsgegenstand der Zoologie, könnte es sein, dass eine Versuchsperson zufällig mit genau diesem Thema (wissenschaftlich) vertraut ist und die vielzitierten Quellen sowie bekannte Forschende auf diesem Gebiet oder angrenzender Gebiete in der Zoologie kennt. Dies könnte dazu führen, dass die manipulierten Popularitätsdaten zu einem Autor oder Werk als falsche Angaben entlarvt würden, wodurch der eigentliche Zweck des Experiments mit hoher Wahrscheinlichkeit korrekt von der Versuchsperson

vermutet würde. In diesem Fall würde sie sich möglicherweise bewusst anders verhalten als die anderen Versuchspersonen, denen diese Hinweise auf das Ziel der Befragung entgehen, und es könnte ein Good-subject-Effekt auftreten. Somit wäre die interne Validität des Experiments gefährdet. Zudem wäre die Erstellung der Surrogate mit der Schwierigkeit verbunden, Quellen zu finden, deren Beurteilung hinsichtlich ihrer Güte und Angemessenheit ohne entsprechende Fachkenntnisse nur mit einem weitaus höheren Aufwand (z. B. Einholgen von Experten-Feedback) zu realisieren gewesen wäre. Als Informationswissenschaftlerin kann die Autorin entsprechende Quellen zu informationswissenschaftlichen Themen besser bewerten als beispielsweise zoologische oder astrophysikalische Quellen.

Zusammengefasst basiert die Themenauswahl auf dieser Anforderung: Die Versuchspersonen sollten Informationsbedürfnisse zu Themen erhalten, die sie zwar kennen, um die Suchergebnisse verstehen und bewerten zu können, derartige Themen aber nicht im Rahmen ihrer Forschung behandeln, um die Gefahr einer möglichen Identifikation der manipulierten Zitations- und Downloadzahlen bestmöglich ausschließen zu können. Die Themen wurden im Zusammenhang mit der Herkunft der Dokumente, die als Vorlage für das Stimulusmaterial in Form der zu bewertenden Surrogate dienten (vgl. Abschnitt 4.2.2), abgeleitet; wenn nicht genügend Dokumente zu einem potenziell geeigneten Thema gefunden werden konnten, wurde jenes Thema nicht weiterverfolgt.

Für den Pretest wurden zunächst Beschreibungstexte und Surrogate zu neun informationswissenschaftlichen Themen entwickelt:

1. Altmetrics
2. Peer Review
3. Wikipedia
4. Data Citation
5. Open Access
6. Google
7. Information Literacy & Gaming
8. Visual Information Seeking
9. Scholarly Communication

Nach der Anpassung des experimentellen Designs auf Basis der Erkenntnisse aus dem Pretest wurden schließlich die drei Themen ausgewählt, die von der Autorin

(auch aufgrund der Rückmeldungen aus dem Pretest) als am besten geeignet und gut verständlich erachtet wurden: Altmetrics, Peer Review, Wikipedia.[28]

4.2.1.3 Formulierung der Aufgaben

Die Entwicklung der Beschreibungstexte der Informationsbedürfnisse erfolgte nach einem zuvor festgelegten Schema. Zunächst sollte der inhaltliche Rahmen, in dem sich das Thema befindet, eingegrenzt und die Motivation hinter dem Bedürfnis verdeutlich werden. Anschließend wurde der Wunsch, nach welchen Information gesucht wird, beschrieben. Dieser stellt das Informationsbedürfnis dar. Die Beschreibungstexte (Vignetten) bestehen aus zwei Teilen: (1) dem Kontext und (2) dem expliziten Informationsbedürfnis. Die Überleitung zu den Suchergebnissen erfolgte mit der Aufforderung zur Bewertung, deren Wortlaut für alle drei Aufgaben identisch ist. Um eventuellen Unklarheiten bezüglich der Aufgabeninstruktion vorzubeugen, wurde das Informationsbedürfnis explizit als solches gekennzeichnet. Aufgabe 1 zum Thema Altmetrics beinhaltete den folgenden Text:

> Soziale Medien wie Twitter und Facebook sind aus dem Internet, wie wir es heute kennen, nicht mehr wegzudenken. Auch in der wissenschaftlichen Kommunikation werden soziale Medien genutzt. Vor diesem Hintergrund wurde vor einigen Jahren die Forderung nach neuen, alternativen Forschungsindikatoren (neben u.a. Publikationen und Drittmitteln) basierend auf Aktivitäten in sozialen Medien laut.
> *Ihr Informationsbedürfnis:*
> *Sie möchten herausfinden, inwieweit diese altmetrics (alternative metrics) als Forschungsindikator geeignet sind.*

Die Aufgabe 2 zum Thema Peer Review wurde wie folgt formuliert:

> Qualität in der Wissenschaft soll durch Peer Review gesichert werden. Die Diskussion über die Vor- und Nachteile von Peer Review ist nicht neu. Menschen werden in ihrer

[28] Die Aufgabenbeschreibungen und die dazugehörigen Surrogate, wie im Fragebogen enthalten, sind zu finden im Anhang 2 im elektronischen Zusatzmaterial.

> Urteilsfindung oft unbewusst beeinflusst (kognitive Verzerrung), was im Peer Review-Prozess dazu führen kann, dass Gutachten als ungerecht wahrgenommen werden.
> *Ihr Informationsbedürfnis:*
> *Sie möchten herausfinden, welche Arten von Verzerrungen (Bias) im Peer Review-Prozess auftreten (können) bzw. wie mit diesen Verzerrungen umgegangen werden kann.*

Die Aufgabe 3 zu dem Thema Wikipedia wurde folgendermaßen beschrieben:

> Viele Menschen nutzen die Online-Enzyklopädie Wikipedia – die deutschsprachige Webseite wird eigenen Angaben zufolge täglich Millionen Mal aufgerufen. Trotz ihrer Beliebtheit wird Wikipedia im Bildungskontext und im Hochschulbereich gemeinhin nicht als zitierfähige Informationsquelle erachtet, da Zweifel an der Güte bzw. Qualität von Wikipedia-Artikeln bestehen.
> *Ihr Informationsbedürfnis:*
> *Sie möchten herausfinden, ob diese Zweifel in Hinblick auf Wikipedia und Lehre berechtigt sind.*

Die Instruktion zur Bewertung folgte jeweils direkt im Anschluss an die Beschreibungen:

> Auf Basis dieser Beschreibung wurde eine Suchanfrage formuliert, die die nachfolgenden Suchergebnisse erzielte. Bitte beurteilen Sie, für wie nützlich Sie jedes Suchergebnis zur Befriedigung des Informationsbedürfnisses halten! (Zur Erinnerung: Die Reihenfolge der Suchergebnisse ist rein zufällig!)

Der Hinweis auf die zufällige Reihenfolge sollte die Versuchspersonen motivieren, bewusst nicht dem typischen Verhalten auf Ergebnisseiten bei der Websuche – das Bevorzugen der ersten drei Treffer – zu folgen, sondern die

Möglichkeiten des Hochscrollens und Anpassens von Bewertungen zu nutzen. Ein solcher Hinweis ist zwar in der traditionellen (I)IR-Forschung nicht üblich, da mit diesem Experiment jedoch weder ein Ranking-Algorithmus noch ein IR-System in irgendeiner Weise evaluiert wurde, sondern Kenntnisse über das Verhalten informationssuchender Personen bezüglich der Anwendung von Kriterien bei der Relevanzbewertung erzielt werden sollen, steht ein derartiger Hinweis nicht im Konflikt mit dem Zweck des Experiments.

Die Informationsbedürfnisse wurden formuliert mit dem Ziel, sich inhaltlich mit den dazugehörigen Surrogaten zu decken. Die Anforderung bestand darin, bei der Auswahl der Surrogate auf eine hohe thematische Relevanz auf der Basis der Aboutness (vgl. Abschnitt 2.1.1) der Surrogate zu achten, um die Bewertungen durch die Versuchspersonen möglichst auf die manipulierten Zitations- und Downloadzahlen zurückführen und andere Faktoren ausschließen zu können. Neben den Variationen der drei UVn sollten alle anderen potenziellen Einflussfaktoren, die jedoch nicht als separate UV manipuliert und untersucht wurden, wie beispielsweise das Erscheinungsjahr und der Zeitschriftentitel, als weitgehend homogen wahrgenommen werden. Demzufolge wurden zunächst die Quellen für die Erstellung der Surrogate ausgewählt und anschließend die Informationsbedürfnisse formuliert.

4.2.2 Erstellung der Surrogate als Bewertungsgegenstand

Die im Rahmen des Experiments zu bewertenden Surrogate basieren auf realen Publikationen, die anhand von Überblicksartikeln (Literaturschauen) zu ausgewählten Themen (vgl. Abschnitt 4.2.1) mittels dem Verfolgen von zitierten und zitierenden Quellen, also dem *backward chaining* und *forward chaining* – einer üblichen Suchstrategie in Information Retrieval-Systemen (Ellis, 1989), die ebenfalls für die Literaturschau der vorliegenden Arbeit (vgl. Kapitel 2) – identifiziert wurden. Diese Vorgehensweise wurde gewählt, weil davon auszugehen ist, dass Quellen in Überblicksartikeln von den Autorinnen und Autoren aufgrund ihrer thematischen Relevanz zitiert werden, d. h. die Bewertung dieser Quellen hinsichtlich ihrer thematischen Relevanz ist bereits durch Experten erfolgt, die sich mit dem Inhalt der zitierten Dokumente auseinandergesetzt (*evaluative judgments*) und deren Eignung, als Quelle in den entsprechenden Überblicksartikel mit aufgenommen zu werden, festgestellt hatten.

Ein alternatives Vorgehen wäre die Auswahl geeigneter Quellen aus Suchergebnislisten nach Eingabe einer Suchanfrage in einem wissenschaftlichen Suchsystem (*predictive judgments*). Es erschien jedoch sinnvoller, auf die bereits

als thematisch relevant zu erachtenden Quellen in den Überblicksartikeln zurückzugreifen.

Unabhängig von dem Vorgehen bei der Quellenauswahl wurde auf das Einbinden einer Suchanfrage in die Aufgabenbeschreibung verzichtet, weil die Vermutung bestand, dass die Versuchspersonen eher auf mit der Suchanfrage übereinstimmende Begriffe im Titel und Abstract achten würden als auf die manipulierten Popularitätsdaten. Es erschien für das Erreichen valider Ergebnisse sinnvoll, diese potenzielle Ablenkungsursache auszuschließen.

Für die drei ausgewählten Themen dienten die folgenden Artikel als Quellen:

(1) Altmetrics:
Sugimoto, C. R., Work, S., Larivière, V., & Haustein, S. (2017). Scholarly use of social media and altmetrics: A review of the literature. Journal of the Association for Information Science and Technology, 68(9), 2037–2062. http://doi.org/10.1002/asi.23833

(2) Peer Review:
Lee, C. J., Sugimoto, C. R., Zhang, G., & Cronin, B. (2013). Bias in peer review. Journal of the American Society for Information Science and Technology, 64(1), 2–17. http://doi.org/10.1002/asi.22784

(3) Wikipedia:
Mesgari, M., Okoli, C., Mehdi, M., Nielsen, F. Å., & Lanamäki, A. (2015). "The sum of all human knowledge": A systematic review of scholarly research on the content of Wikipedia. Journal of the Association for Information Science and Technology, 66(2), 219–245. http://doi.org/10.1002/asi.23172
Okoli, C., Mehdi, M., Mesgari, M., Nielsen, F. Å., & Lanamäki, A. (2014). Wikipedia in the eyes of its beholders: A systematic review of scholarly research on Wikipedia readers and readership. Journal of the Association for Information Science and Technology, 65(12), 2381–2403. http://doi.org/10.1002/asi.23162

Die Auswahl geeigneter Publikationen aus den Überblicksartikeln erfolgte anhand von sechs Kriterien, die zum Teil erst während des Auswahlprozesses aufgestellt wurden:

1. Dokumenttyp: Es sollen ausschließlich Forschungsartikel in wissenschaftlichen Zeitschriften oder Konferenzbänden aus Gründen der Vergleichbarkeit und des Vorhandenseins gleichwertiger Abstracts ausgewählt werden, d. h. keine Monografien, Reports, graue Literatur, Abschlussarbeiten, Workshop-Beiträge oder Poster auf Konferenzen.

2. Quelle: Die Zeitschrift bzw. der Konferenzband soll eine informationswissenschaftliche Quelle sein.
3. Erscheinungsjahr: Die Publikation soll in den Jahren 2010 bis 2016 erschienen sein, was zum Zeitpunkt der Entwicklung der Studie als jung genug erachtet wurde, um als aktuell zu gelten, und als alt genug, um je nach experimenteller Bedingung hohe Download- bzw. Zitationszahlen angesammelt haben zu können; dabei zählt die Angabe der gedruckten Quelle, nicht das Datum eines eventuell vorab online veröffentlichten Dokumentes.
4. Autorenschaft: Es sollen keine Veröffentlichungen von Organisationen, Arbeitsgruppen oder ähnlichen ausgewählt werden, um autorenbezogene Zitationsangaben zu ermöglichen.
5. Autoren: Innerhalb eines Tasks, also innerhalb einer Liste von Surrogaten, soll jeder Autorenname nur einmal vorkommen, um die Wirkung der manipulierten Zitationsangaben nicht zu gefährden; ggf. ist der jüngere Aufsatz zu wählen.
6. Titel: Es sollen keine Quellen mit fachspezifischen Abkürzungen im Titel, die nicht aufgelöst sind, ausgewählt werden.

Da die Überblicksartikel in internationalen, englischsprachigen Publikationen erschienen sind, ist auch die Auswahl der Quellen auf Zeitschriften- und Konferenzbeiträge in englischer Sprache begrenzt. Englisch gilt als Wissenschaftssprache und so war davon auszugehen, dass die Versuchspersonen keine Schwierigkeiten mit dem Verstehen der Surrogate haben würden.

Für die Erstellung der 27 Surrogate wurde eine Vorlage entwickelt, die sich am Design der Suchergebnisdarstellung in Google Scholar orientiert, zum Beispiel bezüglich der blauen und grünen Schriftfarbe; die englischsprachigen Bezeichnungen der Popularitätsdaten und der Systemfunktionalitäten (Hinzufügen zu einer Merkliste, Zitieren und Suche nach ähnlichen Dokumenten) suggerieren die Oberfläche eines internationalen wissenschaftlichen Suchsystems (vgl. Abbildung 4.5). In diese Vorlage wurden jeweils die Metadaten der ausgewählten Publikationen eingefügt. Dafür wurden die in Tabelle 4.4 gelisteten Werte der UV-Ausprägungen zu den einzelnen Surrogaten anhand der in Tabelle 4.7 aufgezeigten Reihenfolge manuell zugewiesen. Zu beachten ist hierbei, dass die Surrogate alle dieselbe, angemessene Länge aufweisen sollten und die Einbindung des vollständigen Abstracts daher nicht möglich war. Zudem konnte jedes Werk nur die Angaben zu einem Autor erhalten; aus diesem Grund wurde bei Publikationen mit mehr als einem Autor oder einer Autorin der erstgenannte Name verwendet unter Berücksichtigung des 5. Auswahlkriteriums, womit ein mehrmaliges Auftreten eines Namens ausgeschlossen werden sollte.

Schließlich wurden die Surrogate als einzelne Bilddateien abgespeichert, wodurch eine randomisierte Präsentation im Fragebogen erfolgen konnte (vgl. Abschnitt 4.2.3). Abbildung 4.6 zeigt exemplarisch das erste Surrogat der ersten Aufgabe *Altmetrics* in der experimentellen Bedingung, in der die Zahl der Downloads hoch, die Zitationszahl des Werkes nicht bekannt und die Anzahl der Zitationen des Autors ebenfalls hoch ist. Der vollständige Fragebogen mit allen Surrogaten ist im Anhang 2 im elektronischen Zusatzmaterial enthalten.

<Titel> Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod eirmod tempor invidunt ut labore et dolore

<Autor> • <Erscheinungsjahr> • <Quelle>

<Abstract> Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem

Downloads: Xxxxxx Citations: Xxxxxx Author's citations: Xxxxxx

+ Add to list „" Cite ≈ Similar works

Abbildung 4.5 Surrogat-Template

Do altmetrics follow the crowd or does the crowd follow altmetrics?

Alhoori, H. • 2014 • Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries

Changes are occurring in scholarly communication as scientific discourse and research activities spread across various social media platforms. In this paper, we study altmetrics on the article and journal levels, investigating whether the online attention received by research articles is related to scholarly impact or may be due to other factors. We define a new metric, Journal Social Impact (JSI), based on eleven data sources: CiteULike, Mendeley, F1000, blogs, Twitter, Facebook, mainstream news outlets, Google Plus, Pinterest, Reddit, and sites ...

Downloads: 5490 Citations: n.a. Author's citations: 2975

+ Add to list „" Cite ≈ Similar works

Abbildung 4.6 Beispiel-Surrogat Nr. 1 aus Aufgabe 1

4.2.3 Konstruktion des Online-Fragebogens

Der Fragebogen wurde mit der Software *EFS Survey* entwickelt. EFS (Enterprise Feedback Suite) Survey ist eine webbasierte Softwarelösung in dem akademischen Programm Unipark der Questback GmbH, mit deren Hilfe sich Online-Befragungen konzipieren, organisieren, durchführen und bis zu einem gewissen Grad auch auswerten lassen. Das Tool ersetzt jedoch nicht spezielle Statistik-Software wie SPSS, mit deren Hilfe sich komplexe statistische Verfahren anwenden lassen. EFS Survey wurde für die Konstruktion des Fragebogens und die Erhebung der Daten verwendet, weil diese Software vielfältige Funktionen bietet, die insbesondere in Hinblick auf die Anforderungen des entwickelten experimentellen Untersuchungsdesigns erforderlich sind, wie beispielsweise die Randomisierung der Fragebogenseiten und der einzelnen Items, ohne zusätzlichen Programmieraufwand zu betreiben. Des Weiteren sind Hilfestellungen durch Tutorials, den Support und Fragen und Antworten im Community-Bereich gegeben. Zudem ist es möglich, die Ergebnisdaten in unterschiedliche Dateiformate zu exportieren, sodass ein Bearbeiten und Auswerten der Daten mit Statistiksoftware wie SPSS problemlos erfolgen kann.

Der vollständige Fragebogen jeweils vor der Freischaltung in der Editor-Ansicht sowie nach der Freischaltung zur Teilnahme ist abgebildet in Anhang 2 im elektronischen Zusatzmaterial und online im Open Science Framework (OSF)[29] abrufbar; ebenso im OSF sind das aus EFS Survey exportierte Codebuch mit den Variablennamen und Item-Eigenschaften.

[29] Siehe https://doi.org/10.17605/OSF.IO/NTWQD.

Dem Fragebogen voran gestellt war das Einwilligungsformular mit Informationen über die Studie und den Kontaktdaten der Autorin als Ansprechperson. Hervorzuheben ist an dieser Stelle die Formulierung des Studienziels. Um die interne Validität des Experiments nicht zu gefährden, ist es notwendig, die Teilnehmenden nicht über dessen wahren Zweck oder zu prüfende Hypothesen aufzuklären. Daher wurde als Untersuchungsziel die Erforschung der Nutzung wissenschaftlicher Suchsysteme genannt. Eine derartige Täuschung widerspricht zwar dem forschungsethischen Prinzip der bewussten Einwilligung, ist jedoch aufgrund der „methodischen Alternativlosigkeit einerseits sowie des Ausschlusses von Nachteilen für die Probanden andererseits“ (Döring & Bortz, 2016, S. 126) als Begründung zulässig.

Für die Bearbeitung des Fragebogens waren als Zeitaufwand 20 bis 30 Minuten im Einwilligungsformular genannt; dieser zeitliche Rahmen wurde anhand der auf den Pretest-Ergebnissen beruhenden Design-Anpassungen geschätzt und erschien in Relation zur Bearbeitungsdauer von den ursprünglich entwickelten neun Aufgaben realistisch. Die Informationen in dem Einwilligungsformular sahen alle Interessierten, bevor sie der Teilnahme ausdrücklich zustimmten oder eine solche ablehnten.

Nach Zustimmung startete die Befragung, die in insgesamt fünf Teile gegliedert ist: Demografische Angaben, Vorabfragebogen, Experiment, Anschlussfragebogen und Abschluss. Diese sind in Abbildung 4.7 auf Basis der Abbildung 4.2 (vgl. Abschnitt 4.1.5) schematisch dargestellt und werden nachfolgend beschrieben.

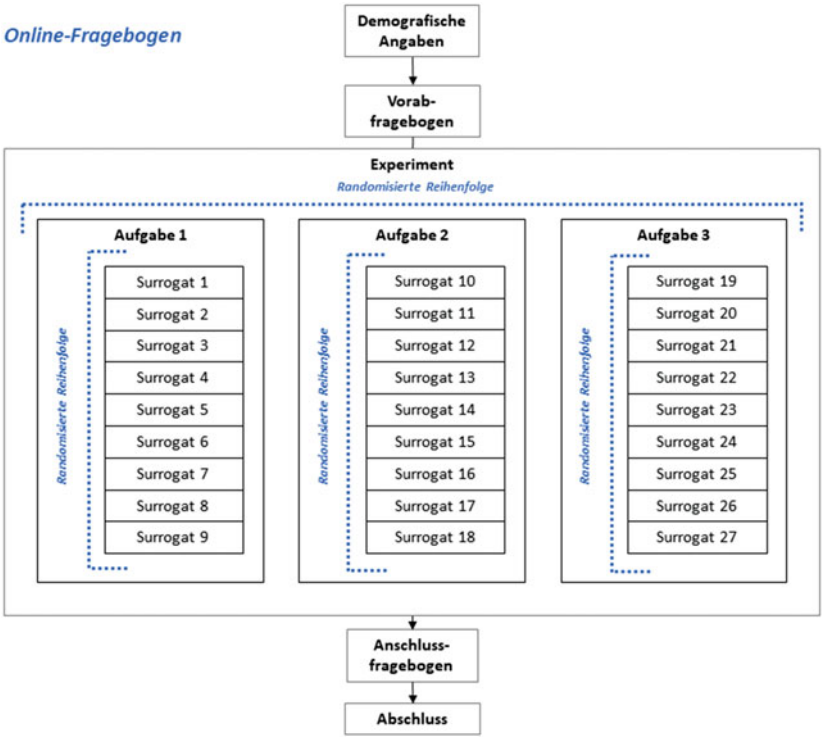

Abbildung 4.7 Schematischer Aufbau des Online-Fragebogens mit dem Experiment als Hauptteil

4.2.3.1 Demografische Angaben

Neben der Erhebung von Geschlecht[30] und Alter dienten die Fragen nach der Muttersprache und dem Bildungsabschluss als Filter für die Teilnahme an der Befragung. Die Personen, die angaben, nicht mindestens ein gutes Verständnis der deutschen Sprache oder nicht mindestens einen Abschluss als Master, Magister oder Diplom zu besitzen, wurden zu einer Abbruchseite geleitet und von

[30] Seit Ende 2018 besteht in Deutschland die Möglichkeit, in das Geburtenregister die Eintragung „divers“ vorzunehmen. Diesem nicht-binären Ansatz von Geschlechtsidentität sollte auch bei diesem Item Rechnung getragen werden, weshalb die Fragestellung nicht explizit auf das biologischen Geschlecht abzielt und als Antwortmöglichkeit neben weiblich und männlich auch divers sowie eine Verweigerung dieser Angabe erlaubt (Döring, 2013).

der weiteren Teilnahme ausgeschlossen. Obwohl es sich bei den zu bewertenden Items um Surrogate in englischer Sprache handelte, waren gute Sprachkenntnisse in Deutsch erforderlich, um die Kontextbeschreibungen und die Informationsbedürfnisse zu verstehen. Die Fragen nach dem aktuellen Status der Teilnehmenden, der Zugehörigkeit zur wissenschaftlichen Fachdisziplin und Art der Einrichtung (Universität, Fachhochschule oder ähnliche) zielten auf ein besseres Verständnis von der Zusammensetzung der Stichprobe und sollten in Teilen der Evaluierung der verfolgten Maßnahmen zur Probandengewinnung dienen. Zudem können mithilfe dieser Angaben die Daten der Versuchspersonen gruppenweise betrachtet werden.

4.2.3.2 Vorabfragebogen

Dieser Teil sollte hauptsächlich die Teilnehmenden auf den Kontext der nachfolgenden Aufgaben vorbereiten und die Behauptung über das vermeintliche Studienziel untermauern. Die Frage nach der Häufigkeit der Nutzung ausgewählter, fachübergreifender akademischer Suchsysteme erlaubt, Erkenntnisse über den Erfahrungsumfang der Versuchspersonen zu gewinnen[31]. Zusätzlich bestand die Möglichkeit, fachspezifische Recherchetools mittels Freitextfeld und deren Nutzungshäufigkeiten anzugeben. Wie häufig bei der Interaktion mit akademischen Suchsystemen auch englischsprachige Quellen berücksichtigt werden, kann Erkenntnisse über die Erfahrungsstufe der jeweiligen Versuchsperson liefern. Zu erwarten wäre, dass ein Großteil der Teilnehmenden diese Frage mit der höchsten Stufe der Skala (5 – immer) beantwortet, wobei sprachliche Präferenzen auch in Abhängigkeit mit der Fachdisziplin oder dem Forschungsgebiet zu betrachten wären. Die Antworten in diesem Teil liefern weitere Informationen über die

[31] In vielen IIR-Studien werden die Teilnehmenden vor dem Bearbeiten von Suchaufgaben nach ihrer Einschätzung der Suchfähigkeit und ihren Erfahrungen mit Suchsystemen befragt. Bisher ist das von Bailey (2017) entwickelte Instrument zur Erfassung der Online-Suchfähigkeit (*search expertise*) das einzige für diese Art von Untersuchung infrage kommende, das nach testtheoretischen Gütekriterien im Rahmen der Fragebogenkonstruktion entwickelt und evaluiert wurde (Bailey & Kelly, 2016). Nichtsdestotrotz konnten die Items aufgrund unterschiedlicher Zielstellungen weder vollständig noch exakt übernommen werden und dienten daher vorwiegend der Inspiration. So sind beispielsweise Fragen zur Selbsteinschätzung der Suchfähigkeit im vorliegenden Experiment nicht von Bedeutung und wurden in Hinblick auf das Gütekriterium der Testökonomie (Döring & Bortz, 2016, S. 449) nicht integriert.

Teilnehmenden und können für weiterführende, explorative Analysen[32] genutzt werden.

4.2.3.3 Experiment

Die Bearbeitung der drei Aufgaben, die in der Bewertung der jeweils neun Surrogate bestand, stellte den Hauptteil der Befragung dar (vgl. Abschnitt 4.1.5). Vorab lasen die Versuchspersonen eine ausführliche Instruktion und sahen ein exemplarisches Surrogat, dessen Aufbau zusätzlich erläutert war. Die Erklärung zur Verwendung der Schiebereglerskala erfolgte ebenfalls anhand von Screenshots und es bestand die Möglichkeit des Ausprobierens. Im Anschluss wurden die Aufgaben jeweils in randomisierter Reihenfolge präsentiert. Zusätzlich erhielten die Teilnehmenden vor Bearbeitung der ersten Aufgabe den ausdrücklichen Hinweis, dass die Reihenfolge der einzelnen Suchergebnisse keinem Ranking folgt, sondern rein zufällig entsteht. Dieser Hinweis wiederholte sich im Anschluss an die Präsentation der jeweiligen Informationsbedürfnisse.

Die Schiebereglerskala, mit der die Versuchspersonen ihre Bewertung kennzeichneten, war im Ausgangszustand ohne den Regler dargestellt, der sich erst durch Klick mit der linken Maustaste auf die Skala (dargestellt als hellgraue Gerade) zeigte (Abbildung 4.8), um die Versuchspersonen nicht mit einer Voreinstellung in ihrer Beurteilung zu beeinflussen (vgl. Abschnitt 4.1.2). Sowohl in der Druckversion des aus EFS Survey exportierten Fragebogens als auch im Codebuch werden die Items leider ohne Schiebereglerskala angezeigt.

[32] Derartige Analysen sind nicht Gegenstand der vorliegenden Arbeit, da sie keinen wesentlichen Beitrag zur Beantwortung der Forschungsfragen leisten; die erhobenen Angaben zu den beschriebenen Bewertungskriterien bieten jedoch eine umfassende Datengrundlage für künftige Untersuchungen nicht nur im Kontext der informationswissenschaftlichen Relevanzforschung, sondern können ebenfalls Einblicke in das Informationssuchverhalten bzw. *Information Searching Behaviour* von Forschenden im Zusammenhang mit akademischen Suchsystemen geben (vgl. Abschnitt 5.3).

Abbildung 4.8 Schiebereglerskala im Ursprungszustand (oben) und nach dem Anklicken (unten)

4.2.3.4 Anschlussfragebogen

Dieser Teil umfasste eine Freitextabfrage über die Kriterien, anhand derer die Versuchspersonen vermuteten, die Bewertung der Surrogate vorgenommen zu haben sowie ein letztes Item mit einer 5-stufien Skala, das nach der Wichtigkeit der drei Arten von Popularitätsdaten (Anzahl der Downloads, Anzahl der Zitationen eines Werkes, Anzahl der Zitationen eines Autors) fragte. Beide Items erschienen in Hinblick auf eine mögliche Diskrepanz der eigenen Wahrnehmung der Versuchspersonen gegenüber ihren tatsächlichen Beurteilungen für eine nähere Betrachtung und ebenfalls für weiterführende, explorative Analysen interessant. Da insbesondere die Abfrage der Wichtigkeit von Zitations- und Downloadzahlen als Bestandteil von Suchergebnissen bei den Teilnehmenden korrekte Annahmen über den wahren Zweck der Befragung zulässt, konnten diese Fragen nur im Anschluss an das Experiment gestellt werden.

4.2.3.5 Abschluss

Vor dem Beenden der Befragung hatten die Teilnehmenden die Möglichkeit über ein Freitextfeld Anmerkungen zum bisherigen Befragungsverlauf einzugeben. Anschließend wurde über den wahren Zweck des Experiments aufgeklärt. Weil die Täuschung über das tatsächliche Studienziel dem forschungsethischen Prinzip der bewussten Einwilligung widerspricht, ist die Aufklärung am Ende zwingend erforderlich und den Teilnehmenden muss die Möglichkeit gegeben werden, nach der Aufklärung ihre Einwilligung zurückzuziehen (Döring & Bortz, 2016, S. 126). Diese war mit dem freiwilligen Selbstausschluss gegeben, mit dem Hinweis, dass eine Zurücknahme der Einwilligung zur Verwendung der erhobenen Daten keine Auswirkung auf die Verlosung hat. Dadurch sollte auch denjenigen Teilnehmenden, die vorrangig wegen der Verlosung den Fragebogen bearbeiteten, die Chance geboten werden, ihre eventuell nicht gewissenhaft getätigten Eingaben

von der Auswertung auszuschließen; dabei wurde auf die Ehrlichkeit der Teilnehmenden gesetzt. Zuletzt konnten die Versuchspersonen ihre E-Mail-Adresse hinterlassen, um an der Gutscheinverlosung teilzunehmen.

Mit dem Erreichen der Endseite war die Befragung abgeschlossen und die Eingaben der Teilnehmenden gespeichert. Diese letzte Seite enthielt neben einer Danksagung nochmals die Kontaktdaten der Autorin, um den Teilnehmenden die Möglichkeit zu geben, Feedback zu geben oder Nachfragen zu stellen, nachdem sie über den Zweck der Befragung aufgeklärt worden waren, da das Freitextfeld für eventuelle Anmerkungen bereits vor der Aufklärung zum Zweck der Studie gezeigt wurde.

4.2.4 Berechnung des erforderlichen Stichprobenumfangs

Die Berechnung des optimalen Stichprobenumfangs vor der Durchführung der Datenerhebung (a priori) ist für einen aussagekräftigen Signifikanztest bei der späteren Datenauswertung von besonderer Bedeutung (Döring & Bortz, 2016, S. 815). Die optimale Stichprobengröße setzt sich aus den drei Parametern Effektgröße, Signifikanzniveau und Teststärke zusammen.

Die Effektgröße ist der Parameter, der auf der Basis von Ergebnissen und ermittelter Effektgrößen aus inhaltlich vergleichbaren Studien geschätzt wird. Wie bei der Formulierung von Hypothesen über Interaktionseffekte der UVn (vgl. Abschnitt 4.1.3) ist dies für das vorliegende Experiment aufgrund fehlender Referenzstudien nicht möglich. Bei der erwarteten Effektgröße wird allgemein der klassischen Einteilung nach Cohen gefolgt (Döring & Bortz, 2016, S. 820). So lassen sich kleine, mittlere und große Effekte unterscheiden. Eine Effektstärke von $f = 0.2$ kann nach Cohen als klein interpretiert werden, bei $f = 0.25$ ist bereits von einem mittleren Effekt auszugehen (Lenhard & Lenhard, 2016). Um im Rahmen des hier beschriebenen Experiments auch kleine Effekte aufdecken zu können, wird in der Berechnung der Parameter f auf 0.2 gesetzt.

Traditionell wird bei quantitativen Studien von einem Signifikanzniveau $\alpha = 0.05$ und einer Teststärke $1 - \beta = 0.80$ ausgegangen (Döring & Bortz, 2016, S. 841); ersteres wird für die Berechnung beibehalten, die Teststärke jedoch folgt dem Wert, der im für die Berechnung verwendeten Tool G*Power[33] (Faul et al., 2007) auf 0.95 voreingestellt ist. Abbildung 4.9 zeigt einen Screenshot

[33] Informationen zur Software und Downloadmöglichkeit sind zu finden unter http://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower.html (letzter Zugriff: 09.10.2019).

der Eingabeoberfläche von G*Power mit den ausgewählten Parametern und eingesetzten Werten der A-priori-Analyse: Als statistischer Test wurde die Art der Varianzanalyse (*Analysis of Variance*, ANOVA) ausgewählt, die das Prüfen von Haupteffekten und Interaktionen in einem mehrfaktoriellen Design zulässt. Die dem statistischen Modell zugrunde liegende Methode der Mehrebenenanalyse (auch Hierarchisches Lineares Modell) (vgl. Abschnitt 4.3), kann in G*Power nicht ausgewählt werden; da diese letztlich auch als eine Form der Varianzanalyse gesehen werden kann (Hoffman & Rovine, 2007, S. 103), wird für die Berechnung des optimalen Stichprobenumfangs die genannte ANOVA zugrunde gelegt.

Der *Numerator df* setzt sich laut Handbuch[34] aus der Anzahl der Faktoren jeweils um 1 subtrahiert und dann multipliziert zusammen, für das vorliegende $3 \times 3 \times 3$-Design ergibt sich demzufolge der Wert 8 ($2 \times 2 \times 2$). Das Feld *Number of groups* verlangt die Anzahl der zu untersuchenden Teilstichproben, also der experimentellen Bedingungen.

Die optimale Stichprobengröße beträgt $n = 577$. Für die im folgenden Abschnitt beschriebene Probandenakquise war von einer gewissen Abbruchrate auszugehen, daher war zu berücksichtigen, dass eine weitaus größere Anzahl potenziell Teilnehmender eingeladen werden musste.

4.2.5 Probandenakquise

Die Auswahl möglicher Probandinnen und Probanden folgte keinem probabilistischen Verfahren, wodurch sie eine eingeschränkte Repräsentativität aufweist; stattdessen stellt die Stichprobe eine Gelegenheitsstichprobe dar (Döring & Bortz, 2016, S. 305 ff.), für deren Zustandekommen ein systematisches Vorgehen gewählt wurde, um einer repräsentativen Stichprobe zumindest nahezukommen.

Während die Probandenakquise für den Pretest auf die Wahl einer Hochschule begrenzt war, wurden für die eigentliche Erhebung gruppenweise adressierte E-Mails an Wissenschaftliche Angestellte, Promovierende und Postdocs in verschiedenen Universitäten von der Autorin verschickt. Laut Hochschulkompass[35] gibt es in Deutschland 87 Universitäten in öffentlich-rechtlicher Trägerschaft.

[34] Informationen zu den Feldern und Funktionen des Tools sind zu finden im G*Power Manual für die Version 3.1 von 2017 unter http://www.gpower.hhu.de/fileadmin/redaktion/Fakultaeten/Mathematisch-Naturwissenschaftliche_Fakultaet/Psychologie/AAP/gpower/GPowerManual.pdf (letzter Zugriff: 29.05.2020).

[35] Liste aller Hochschulen in Deutschland herunterladbar via: https://www.hochschulkompass.de/hochschulen/downloads.html.

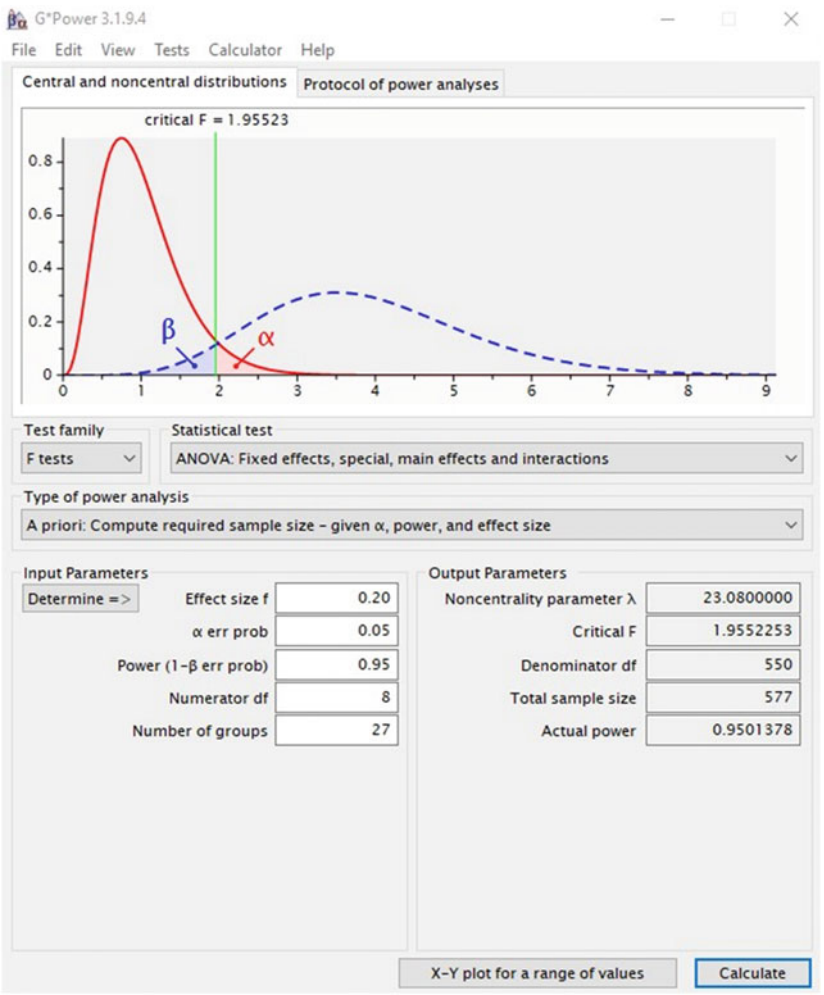

Abbildung 4.9 Berechnung des optimalen Stichprobenumfangs mit G*Power (Screenshot)

Die Reihenfolge, in der das Einholen von E-Mail-Adressen der 87 Universitäten erfolgte, wurde erneut mithilfe des Tools *Research Randomizer* festgelegt. Neben dem Ausschluss von Bibliotheks- und Informationswissenschaftlichen Instituten sowie Einrichtungen mit dem Namen Informations- und Wissensmanagement (aufgrund der Nähe zur Informationswissenschaft) erfolgt die Erfassung der E-Mail-Adressen anhand forschungspragmatischer Kriterien: Um den Aufwand so gering wie möglich zu halten, sollten die E-Mail-Adressen leicht zugänglich sein, also möglichst blockweise kopierbar und ohne zusätzliches Aufrufen der jeweiligen Webseiten von Einzelpersonen. Beispielsweise konnten von der Universität Hamburg insgesamt 1.734 E-Mail-Adressen von Wissenschaftlichen Mitarbeiterinnen und Mitarbeitern relativ mühelos erfasst werden; bei anderen Universitätswebseiten waren die E-Mail-Adressen von Mitarbeitenden vermutlich aus Schutz vor Spam mit Kopierschutzmaßnahmen versehen, auf deren Erfassen in einem solchen Fall verzichtet und stattdessen die nächste Universitätswebseite besucht wurde.

Der Text der E-Mail-Einladung lautete folgendermaßen und wurde mit den jeweils auszutauschenden Elementen (Name des Instituts bzw. des Fachbereichs, der Universität sowie die jeweilige URL als Quelle der E-Mail-Adresse aus Gründen der Transparenz) entsprechend manuell angepasst:

> Betreff: Teilnehmende für Online-Umfrage zur Nutzung wissenschaftlicher Suchsysteme gesucht
>
> Sehr geehrte Wissenschaftlichen Mitarbeiterinnen und Mitarbeiter am <Institut für/Fachbereich XY> der <ABC-Universität>,
>
> für eine groß angelegte Studie zum Thema "Nutzung von wissenschaftlichen Suchsystemen" suchen wir eine hohe Zahl an Teilnehmenden unterschiedlicher Fachrichtungen, die einen **Online-Fragebogen** bearbeiten. Die Studie wird im Rahmen eines Forschungsprojekts an der Hochschule für Angewandte Wissenschaften Hamburg in Kooperation mit der Universität Hildesheim durchgeführt.
>
> Die Bearbeitungszeit des Online-Fragebogens beträgt ca. **20 bis 30 min**. Nach Abschluss der Umfrage haben Sie die Möglichkeit, an einer **Verlosung von insgesamt 444**

Gutscheinen von Amazon.de im Wert von jeweils 10 EUR teilzunehmen!

Weitere Informationen und den Fragebogen finden Sie hier: https://ww3.unipark.de/uc/Studie_Nutzung_wiss_Suchsysteme2019/

Sie erhalten diese E-Mail, weil Ihre Adresse über die Webseite Ihrer Universität (<URL>) manuell erfasst wurde, wobei die Auswahl der Universität zufällig erfolgte und insbesondere Wissenschaftliche Mitarbeiterinnen und Mitarbeiter bzw. Promovierende oder PostDocs für die Teilnahme gesucht werden. Die Teilnahme ist selbstverständlich freiwillig und ohne Angabe von Namen möglich; eine Registrierung ist nicht erforderlich.

Mit freundlichen Grüßen aus Hamburg
Christiane Behnert
--
Christiane Behnert, M.A.
Wissenschaftliche Mitarbeiterin
T +49 40 428 75 3642
christiane.behnert@haw-hamburg.de
HOCHSCHULE FÜR ANGEWANDTE
WISSENSCHAFTEN HAMBURG
Fakultät Design, Medien & Information
Department Information
Finkenau 35 / 22081 Hamburg
haw-hamburg.de
http://searchstudies.org/christiane-behnert/
http://orcid.org/0000-0002-4863-6118

Insgesamt wurden 16.137 E-Mail-Adressen manuell ermittelt und in 1.145 E-Mails[36] an Angehörige von 31 Universitäten in dem Zeitraum vom 6. Juni bis

[36] Die Anzahl der versendeten E-Mails ist wesentlich kleiner als die Anzahl der angeschriebenen Angehörigen, weil eine E-Mail jeweils an mehrere Angehörige eines Instituts oder Fachbereichs adressiert wurde.

12. Juli 2019[37] versendet. Von diesen E-Mail-Adressen waren 545 fehlerhaft oder ungültig und daher unzustellbar; von den 15.592 gültigen Adressen erreichten die Autorin 126 Abwesenheitsinformationen aufgrund von Urlaub oder Mutterschutz- und Elternzeiten. Somit wurden insgesamt 15.466 E-Mail-Empfängerinnen und -Empfänger erreicht.

Die Zahl der Teilnehmenden wurde während der Feldzeit[38] in EFS Survey überwacht, um im Falle eines Nichterreichens der angestrebten Stichprobengröße eine alternative Rekrutierungsstrategie zu verfolgen. Diese bestand darin, die Studie in SurveyCircle[39] zu teilen. SurveyCircly ist eine Online-Crowdsourcing-Plattform für das Anbieten von und Teilnehmen an empirischen Studien, die von Questback empfohlen wird, jedoch nur maximal 100 Teilnehmende pro Studie erlaubt. Diese Strategie zur Probandenakquise musste jedoch nicht umgesetzt werden, da sich zum 12. Juli 2019 abzeichnete, dass eine mehr als ausreichend große Anzahl an Teilnehmenden erreicht werden würde; so hatten bis dato 675 Versuchspersonen den Fragebogen abschließend bearbeitet.

Als Anreiz zur Teilnahme standen 444 Gutscheine für den Online-Shop Amazon.de im Wert von jeweils 10 EUR zur Verfügung, die im Rahmen einer Verlosung an 444 Personen nach Abschluss der Feldzeit per E-Mail versendet wurden. Dieser Wert wurde lediglich als Geste der Dankbarkeit gewählt und ist nicht als eine realistische Aufwandsentschädigung zu betrachten.

Die hier beschriebene Strategie zur Probandenakquise war mit einem hohen Zeit- und Kostenaufwand verbunden, sie ist jedoch aufgrund des erreichten Stichprobenumfangs als erfolgreich zu bewerten.

4.3 Datenaufbereitung und statistische Analyse

In diesem Abschnitt wird das Vorgehen bei der Aufbereitung und der Analyse der Daten beschrieben – für beides wurde die Statistik-Software SPSS von IBM in der Version 25 verwendet.

Die Datenbereinigung und Datenaufbereitung sind in Abschnitt 4.3.1 dokumentiert. Hierzu zählt auch die Codierung der Angaben der Versuchspersonen zu ihrem jeweiligen wissenschaftlichen Fachgebiet bei der Bearbeitung des Vorabfragebogens.

[37] Der Erhebungszeitraum umfasste insgesamt 36 Tage.

[38] Es bestand die Möglichkeit, ab dem Erhalt der E-Mail-Einladung, also ab dem 6. Juni 2019, bis zum 31. Juli 2019, 22.00 Uhr, auf den Online-Fragebogen zuzugreifen.

[39] https://www.surveycircle.com/de/ (letzter Zugriff: 22.05.2020).

Für die statistische Auswertung experimentell erhobener Daten wird in der Psychologie oft auf die Varianzanalyse (auch ANOVA – *Analysis of Variance*) zurückgegriffen (Sedlmeier & Renkewitz, 2018, S. 430). Für das hier berichtete Online-Experiment wurden die Daten nicht mithilfe der Varianzanalyse ausgewertet, sondern es wurde eine Mehrebenenanalyse durchgeführt. Der Grund für die Wahl dieser statistischen Analysemethode und das Vorgehen in SPSS werden in Abschnitt 4.3.2 näher beleuchtet.

4.3.1 Datenaufbereitung

Die erhobenen Rohdaten wurden zunächst mithilfe der automatischen Datenbereinigungsfunktion in EFS Survey von eventuell enthaltenen irrelevanten Werten befreit. Das Ziel der manuellen Datenbereinigung war es, den Datensatz um die Eingaben der Versuchspersonen zu bereinigen, die zum einen den Fragebogen tatsächlich nicht beendet hatten (n = 58), obwohl EFS Survey diese fälschlich als beendet identifizierte, und die zum anderen den Fragebogen zwar beendet, aber aufgrund des gewählten freiwilligen Selbstausschlusses (n = 32) ihre Eingaben nicht für die statistische Analyse freigegeben hatten. Von den insgesamt 717 durch EFS Survey als beendet deklarierte Fragebögen konnten die Daten von 627 Teilnehmenden in die Auswertung einfließen.[40]

Ferner wurden im Rahmen der Verlosung der 444 Amazon-Gutscheine die E-Mail-Adressen der Teilnehmenden entfernt. Wie Tabelle 4.9 zeigt, nahmen an der Verlosung 530 Versuchspersonen teil, 129 hinterließen keine E-Mail-Adresse, wodurch sie nicht an der Verlosung teilnahmen. Der Anteil der Teilnehmenden, die den Selbstausschluss wählten und zugleich an der Verlosung teilnahmen (n = 20) ist zwar relativ gering, darf aber als nicht unerheblich gesehen werden; insofern ist die Entscheidung, die Verlosung unabhängig von einem Selbstausschluss zu erlauben, als richtig zu beurteilen.

[40] Die Teilnahmequote (Rücklaufquote) beträgt 10,73 %; die Anzahl derjenigen Teilnehmenden, die den Fragebogen tatsächlich beendeten, beträgt 659, was einer Beendigungsquote von 4,23 % entspricht, gemessen an der Gesamtzahl der via E-Mail erreichten Personen von 15.466.

Tabelle 4.9 Anteil der Versuchspersonen mit gewähltem Selbstausschluss und Teilnahme an der Verlosung

		Teilnahme an Verlosung		Gesamt
		nein	ja	
Freiwilliger Selbstausschluss	Ja, ich habe alle Fragen sinnvoll beantwortet. Meine Angaben können ausgewertet werden.	117	510	627
	Nein, ich wollte mir nur einmal aus Neugierde die Umfrage anschauen, nehme zum wiederholten Mal teil oder möchte nicht, dass meine Angaben für die Auswertung verwendet werden.	12	20	32
	Gesamt	129	530	659

Im Anschluss an die Datenbereinigung erfolgte die eigentliche Aufbereitung. Hierunter fallen die Codierung der Eingaben in Freitextfelder im Vorabfragebogen. So wurden die Angaben zu den Fachdisziplinen durch die Versuchspersonen auf der Grundlage der Field of Science and Technology Classification der OECD codiert. Diese Klassifikation besitzt insgesamt sechs Hauptklassen und 42 Unterklassen; die Codierung der Angaben beschränkte sich jedoch auf die sechs Hauptklassen, weil diese für eine erste Gruppierung der Ergebnisse nach Fachzugehörigkeit als angemessen erachtet wurden. Die Hauptklassen sind nachfolgend aufgelistet; zum besseren Verständnis der Einordnung der Fachgebiete werden die Unterklassen zu den Hauptklassen ebenfalls aufgezählt (OECD, 2007, S. 6 ff.):

1. **Natural sciences:**
   Mathematics; Computer and information sciences; Physical sciences; Chemical sciences; Earth and related Environmental sciences; Biological sciences (Medical to be 3, and Agricultural to be 4); Other natural sciences
2. **Engineering and technology**
   Civil engineering; Electrical engineering, Electronic engineering, Information engineering, Mechanical engineering; Chemical engineering; Materials engineering; Medical engineering; Environmental engineering; Environmental

biotechnology; Industrial biotechnology; Nano-technology; Other engineering and technologies

3. **Medical and Health sciences**
   Basic medicine; Clinical medicine; Health sciences; Medical biotechnology; Other medical sciences
4. **Agricultural sciences**
   Agriculture, Forestry, and Fisheries; Animal and Dairy science; Veterinary science; Agricultural biotechnology; Other agricultural sciences
5. **Social sciences**
   Psychology; Economics and Business; Educational sciences; Sociology; Law; Political science; Social and economic geography; Media and communications; Other social sciences
6. **Humanities**
   History and Archaeology; Languages and Literature; Philosophy, Ethics and Religion; Arts (arts, history of arts, performing arts, music); Other humanities

Während die überwiegende Anzahl der Angaben eindeutig zu einer Hauptkategorie zuordbar waren, gab es Begriffe, die unscharf oder nicht codierbar waren. Unscharfe Begriffe waren beispielsweise „Energiewirtschaft“ und „Energieeffizienz“ – während ersteres eher dem Bereich Management und Industrie und demzufolge der Hauptklasse (5) zuzurechnen ist, zählt letztgenannter als Bereich des *Environmental engineering* zur Hauptklasse (2). Nicht codierbar waren aufgrund von Unsicherheit Angaben wie „Visualisierung“ oder „Grundlagenforschung“, allerdings auch „Geographie“, da diese generell zwei verschiedenen Hauptklassen zugeordnet werden kann (*physical geography* zählt zu den Naturwissenschaften, *economic geography* zu den Sozialwissenschaften).

Enthielten die Angaben mehrere Begriffe, wurde nach dem Fachgebiet codiert, das vordergründig das intendierte Fachgebiet zu sein schien, zum Beispiel wurde „Veterinärphysiologie, Aminosäuretransport“ als Hauptklasse (4) codiert, da davon auszugehen ist, dass das Thema Aminosäuretransport sich nicht auf den Menschen, sondern auf Tiere oder Pflanzen bezieht. Bei mehreren genannten Fachgebieten, die zu verschiedenen Hauptgruppen gehören, zählte das erstgenannte, beispielsweise wurde „Elektrotechnik, Informatik“ mit einer 2 (für Elektrotechnik) anstelle einer 1 (für Informatik) codiert.

Abschließend erfolgte für die Analyse eine Umstrukturierung der Daten, um diese in eine für die Mehrebenenanalyse mit SPSS erforderliche Form zu bringen. Die Übertragung der im Wide-Format vorliegenden Daten in das Long-Format geschah mithilfe des SPSS-Assistenten für die Datenumstrukturierung, indem

ausgewählte Variablen (Daten spaltenweise) in Fälle (Daten zeilenweise) umgesetzt wurden. Dazu wurde jeweils eine zusätzliche Spalte für jede UV manuell eingefügt und zu jeder Bewertung jeweils die Ausprägung (1 – gering, 2 – hoch, 3 – k.A.) zugeordnet. Das Ergebnis dieser Umstrukturierung besteht darin, dass zu jeder Person die jeweils 27 Bewertungen untereinander gelistet sind, wodurch nun der Datensatz im Long-Format nicht mehr 627 Fälle, sondern 16.929 Fälle beinhaltet. Ferner erfolgte die Umbenennung einiger Variablen in aussagekräftigere Bezeichnungen, wie z. B. die Variablen *a* für Alter und *g* für Geschlecht.

Der vollständige Datensatz steht für Forschungszwecke jeweils im Wide-Format mit den codierten Angaben zur Variable *Fachdisziplin* der 627 Fälle sowie im Long-Format mit den 16.929 Fällen zum Zweck der Forschungstransparenz und zur Nachnutzung im Open Science Framework[41] zur Verfügung, ebenso der vollständig erhobene Rohdatensatz ohne die E-Mail-Adressen der Versuchspersonen.

4.3.2 Mehrebenenanalyse in SPSS

Die aufbereiteten Daten des Experiments wurden einer Mehrebenenanalyse unterzogen, da eine dreifaktorielle Varianzanalyse mit Messwiederholung (RM-ANOVA) aufgrund der Beschaffenheit der Daten nicht infrage kam. Die RM-ANOVA kommt zum Einsatz bei mehreren abhängigen (verbundenen) Stichproben. Abhängige Stichproben liegen dann vor, wenn mehrere Messwerte aus verschiedenen (Teil-)Stichproben voneinander beeinflusst werden und sich diese Abhängigkeiten systematisch über die Stichproben verteilen (Eid et al., 2017, S. 367 ff.). Das ist beispielsweise der Fall bei Messwerten, die von den gleichen Personen unter verschiedenen (experimentellen) Bedingungen, also auf Basis von intraindividuellen Bedingungsvariationen, oder zu unterschiedlichen Zeitpunkten (Messwiederholungen) erhoben wurden. Eine RM-ANOVA berücksichtigt wiederholte Messungen zu unterschiedlichen Zeitpunkten an demselben Objekt bzw. Subjekt, d. h. die Bedingungen, unter denen die Beobachtung am Objekt oder Subjekt vorgenommen werden, sind immer dieselben, nur der Zeitpunkt der Messung ändert sich. Ein Beispiel wäre die Untersuchung der Wirkung eines Medikaments (UV) auf den Menschen (AV) bei einer Experimentalgruppe und einer Kontrollgruppe, wobei den Probanden in der Experimentalgruppe das Medikament (Stufe A der UV) und den Probanden der Kontrollgruppe ein Placebo (Stufe B der UV) verabreicht wird. Werden nun verschiedene Kennzahlen an den

[41] Siehe https://doi.org/10.17605/OSF.IO/NTWQD.

Probanden wiederholt erhoben, geschieht dies unter den gleichen Bedingungen, da es sich immer um dasselbe Medikament bzw. Placebo in derselben Dosierung an derselben Person handelt.

In der vorliegenden Studie sind die Stichproben zwar auch verbunden, da die 27 Relevanzbewertungen jeweils von denselben Personen erhoben wurden. Die Erhebung fand jedoch unter 27 unterschiedlichen experimentellen Bedingungen statt, d. h. es gibt zwar übertragen auf die Logik der Varianzanalyse mit Messwiederholung 27 Messwerte zu jeder Versuchsperson; die Relevanzbewertungen beruhen aber auf 27 verschiedenen Bedingungen. Konkret handelt es sich um 27 intraindividuelle Bedingungsvariationen, wobei die Variable „Zeit" keinen zu untersuchenden Einflussfaktor darstellt. Somit liegen genau genommen keine Messwiederholungen im Sinne einer RM-ANOVA vor (Sedlmeier & Renkewitz, 2018, S. 155–156), sondern hierarchische Datenstrukturen, für deren Auswertung eine Mehrebenenanalyse angemessen ist. Die Logik der Varianzanalyse mit Messwiederholung lässt sich allerdings auf hierarchische Datenstrukturen übertragen (Eid et al., 2017, S. 730); Hoffman & Rovine (2007) bezeichnen die RM-ANOVA als restriktive Version der Mehrebenenanalyse.

Die Mehrebenenanalyse (*multilevel analysis*[42]) stellt eine Form der Regressionsanalyse dar, die zwei oder mehr Ebenen einer Datenstruktur berücksichtigt (Eid u. a., 2017, S. 727). Mit dieser Analysestrategie werden Daten in ein hierarchisches lineares Modell überführt. Die einfachste hierarchische Datenstruktur weist zwei Ebenen auf: Level 1 ist die Mikroebene, auch Individualebene (Richter & Naumann, 2002, S. 155), Level 2 die Makroebene, auch Kontextebene, wobei Level 1 Level 2 untergeordnet ist. In der Literatur wird häufig zur Veranschaulichung der Methode das Beispiel einer Schulklasse vorgebracht: Die Schulklasse wäre ein Merkmal der Ebene 2 und erhobene Merkmale der Schülerinnen und Schüler innerhalb dieser Klasse wären Daten der Ebene 1 (Eid et al., 2017, S. 727 ff.). Übertragen auf die Datenstruktur in dem hier berichteten Experiment stellt eine Versuchsperson die Ebene 2 und die 27 erfassten Bewertungen dieser Person die Daten der Ebene 1 dar (Nezlek et al., 2006, S. 218).

Die Methode der Mehrebenenanalyse wurde bisher selten in IIR-Studien genutzt, ist aber aufgrund verschiedener Vorteile gegenüber Standardverfahren wie der Varianzanalyse auch für IIR-Studien mit experimentellen Designs zu empfehlen (Crescenzi et al., 2016). So können neben den manipulierten unabhängigen Variablen weitere Einflussfaktoren, wie das Suchverhalten der

[42] In der Literatur werden auch folgende alternative Begriffe zur Mehrebenenanalyse bzw. zu Hierarchischen Linearen Modellen genannt: *multilevel (linear) models, mixed effects models, random effects models, random coefficient regression models* und *covariance components models* (Baltes-Götz, 2019).

Teilnehmenden, auf die abhängige Variable berücksichtigt werden und es ist nicht notwendig, Haupt- und Interaktionseffekte separat zwischen den einzelnen Kennzahlen zu suchen – beide können gleichzeitig untersucht werden.

Weitere Vorteile dieser Analysemethode für Daten mit einer hierarchischen Struktur, wie sie durch Within-Subjects-Designs entstehen können, bestehen insbesondere in der größeren Flexibilität, Abhängigkeiten zwischen den beobachteten Messwerten zu berücksichtigen und in der Toleranz fehlender Werte – ob unerwünscht oder bedingt durch das Forschungsdesign (Hoffman & Rovine, 2007). Diese Fehlertoleranz kann für das Design künftiger Experimente sehr nützlich sein, da beispielsweise nicht alle Versuchspersonen immer allen Bedingungen ausgesetzt sein müssen, wodurch die Anzahl der zu bearbeitenden Aufgaben und somit der zeitliche Aufwand für die Teilnehmenden reduziert würde, ohne Einbußen bei der Effektstärke hinzunehmen.

Hierarchische lineare Modelle werden in SPSS als „lineare gemischte Modelle“ bezeichnet und mithilfe der Funktion MIXED (SPSS-Syntax) erstellt. Da nicht von konstanten Abhängigkeiten zwischen den Teilstichproben (Bedingungen) ausgegangen wird, bedeutet dies, dass die Wirkung der Faktoren nicht für alle Versuchspersonen gleich angenommen wird, sondern Faktoren zwischen den Personen verschieden wirken können. Dies zeigt sich in der als diagonal gewählten Kovarianzstruktur (vgl. Tabelle 4.10) im Gegensatz zur identischen Kovarianzstruktur.

Die SPSS-Syntax der Mehrebenenanalyse ist zu finden in Anhang 3 im elektronischen Zusatzmaterial sowie im Open Science Framework[43]. Tabelle 4.10 gibt einen Überblick über die Modelldimension laut SPSS-Ausgabe. Als feste Effekte gelten die unabhängigen Variablen, deren jeweilige Haupteffekte sowie in Kombination untereinander mögliche Interaktionseffekte (vgl. 4.1.3) untersucht wurden. Dabei sind alle Bewertungen (ID_B) aller Personen (ID_P) enthalten, wobei die 27 Bewertungen der jeweils 627 verarbeiteten Fälle als wiederholte Effekte behandelt wurden. Die Korrektur, die zur Kompensation der alpha-Fehlerkumulierung genutzt wurde, ist die Anpassung nach Sidak.

Die Grafiken der Mittelwerte zu den einzelnen unabhängigen Variablen und in Kombination werden mit der MIXED-Funktion in SPSS nicht automatisch erstellt und müssen manuell erzeugt werden. Zur Vermeidung von Fehlern im Zuge einer manuellen Anpassung wurde für die Erzeugung der Grafiken auf die

[43] Siehe https://doi.org/10.17605/OSF.IO/NTWQD.

ANOVA mit Messwiederholung (GLM-Rep[44]-Funktion in SPSS) zurückgegriffen. Dies ist durchführbar, weil die berechneten Mittelwerte die gleichen wie bei der Mehrebenenanalyse sind.

4.4 Ergebnisse des Experiments

In diesem Abschnitt werden die Ergebnisse der Studie vorgestellt, beginnend mit der Beschreibung der demografischen Merkmale der Teilnehmenden (Abschnitt 4.4.1). Anschließend erfolgt die Vorstellung und Erläuterung der Ergebnisse der statistischen Mehrebenenanalyse. Das Ziel der Analyse bestand darin, Haupteffekte der unabhängigen Variablen und mögliche Interaktionseffekte aufzuzeigen und Kovarianzen (Abhängigkeiten zwischen den Bewertungen pro Person) zu berücksichtigen.

In Abschnitt 4.4.2 werden die Haupteffekte, also die Wirkungen der einzelnen unabhängigen Variablen, beschrieben. Jeder Haupteffekt hat einen statistisch signifikanten Einfluss auf die Relevanzbewertung. Allerdings zeigt sich, dass aufgrund der bestehenden Interaktionseffekte (Abschnitt 4.4.3) diese Haupteffekte allein wenig Aussagekraft besitzen, denn „[e]ine Interaktion bedeutet stets, dass der Haupteffekt einer UV nicht über die Stufen der anderen UV generalisiert werden kann“ (Sedlmeier & Renkewitz, 2018, S. 178). Neben der Darstellung der Mittelwerte sind die statistischen Ergebnisse der paarweisen Vergleiche wichtig für die Bestimmung von Richtung und Stärke eines Effekts. Die Effekte zeigen sich jedoch nicht in der erwarteten Richtung, wie im anschließenden Abschnitt 4.5 ausführlich diskutiert wird.

4.4.1 Beschreibung der Stichprobe

Im Zuge der Datenbereinigung und -aufbereitung wurde die Anzahl der Datensätze (n = 627) ermittelt, die für die weitere Analyse verwendet werden können (vgl. Abschnitt 4.3.1). Die Stichprobe besteht aus 291 (46,86 %) Frauen, 329 (52,98 %) Männern, eine Person wählte bei der Frage nach dem Geschlecht die Antwortkategorie divers, sechs Personen gaben ihr Geschlecht nicht an (Abbildung 4.10). Insgesamt gaben 573 (91,39 %) Deutsch als Muttersprache an, 54 (8,61 %) Versuchspersonen gaben an, dass sie mindestens ein gutes

[44] Mit der Funktion *GLM-Rep* werden in SPSS Allgemeine Lineare Modelle mit Messwiederholung bezeichnet.

Tabelle 4.10 SPSS-Ausgabe der Modelldimension

		Anzahl Ausprägungen	Kovarianzstruktur	Anzahl Parameter	Subjektvariablen	Anzahl Fälle
Feste Effekte	Konstanter Term	1		1		
	UV1	3		2		
	UV2	3		2		
	UV3	3		2		
	UV1 * UV2	9		4		
	UV1 * UV3	9		4		
	UV2 * UV3	9		4		
	UV1 * UV2 * UV3	27		8		
Wiederholte Effekte	ID_B	27	Diagonal	27	ID_P	627
Gesamt		91		54		

Verständnis der deutschen Sprache besitzen, unabhängig ihrer Muttersprache (Abbildung 4.11).

Von den 627 Versuchspersonen (VPn) bilden Promotionsstudierende bzw. wissenschaftliche Mitarbeiterinnen und Mitarbeiter mit 448 (71,45 %) den größten Anteil, gefolgt mit deutlichem Abstand von der Gruppe der Postdocs (145; 23,13 %) und der Gruppe der wissenschaftlich Mitarbeitenden ohne Promotionsabsicht (28; 4,47 %); von den sechs Versuchspersonen, die einen anderen Status nannten, gaben zwei VPn „Lehrkraft für besondere Aufgaben“ an, die weiteren Freitextangaben lauten: „Akad. Rat“, „Akademische Oberrätin, Gruppenleiterin“, „Habilitand“, „Studienreferendar, externer Doktorand“ (Abbildung 4.12). Als höchsten Bildungsabschluss nannten 477 (76,08 %) Personen den Master-, Magister- bzw. Diplomabschluss, 150 (23,92 %) gaben den Doktortitel an (Abbildung 4.13).

Zum Zeitpunkt der Datenerhebung gehörten 620 (98,9 %) VPn einer Universität an, 7 VPn gaben an, einer Fachhochschule bzw. Hochschule für Angewandte Wissenschaften anzugehören, wobei alle VPn bis auf eine zugleich einer Universität angehört. Diese Verteilung der Angaben zur Affiliation ist aufgrund der Art der Probandenakquise wenig überraschend, die Überlappung der zeitgleichen Zugehörigkeit einer Universität und einer Fachhochschule lässt sich mit der Durchführung kooperativer Promotionen begründen. Des Weiteren wählten 18 VPn die außeruniversitäre Forschungseinrichtung, 5 Probanden Non-Profit-Organisation, 8 Personen die Kategorie Firma und 15 VPn Selbständigkeit bzw. Freiberuflichkeit aus. Fünf Versuchspersonen nannten jeweils die folgenden sonstigen Einrichtungen: „Gymnasium“, „Institut“, „öffentlicher Dienst“, „Schulpsychologische Beratungsstelle“, „Stipendium“; aufgrund erlaubter Mehrfachnennungen übersteigt die Zahl der Angaben die Anzahl der Teilnehmenden (Abbildung 4.14).

Die Stichprobe setzt sich aus Angehörigen aller sechs Fachdisziplinen nach der verwendeten OECD-Klassifikation (vgl. Abschnitt 4.3.1) zusammen: Den größten Anteil bilden mit 264 (42,1 %) die Naturwissenschaften, gefolgt von den Sozialwissenschaften mit 155 (24,7 %) Versuchspersonen und den Technischen Wissenschaften mit 97 (15,5 %) Versuchspersonen; weniger als ein Drittel der Stichprobe gab einen geisteswissenschaftlichen (48; 7,7 %), humanmedizinischen (27; 4,3 %) oder einen agrarwissenschaftlichen bzw. veterinärmedizinischen (22; 3,5 %) Hintergrund an, die Angaben von 14 (2,2 %) VPn waren nicht codierbar (Abbildung 4.15).

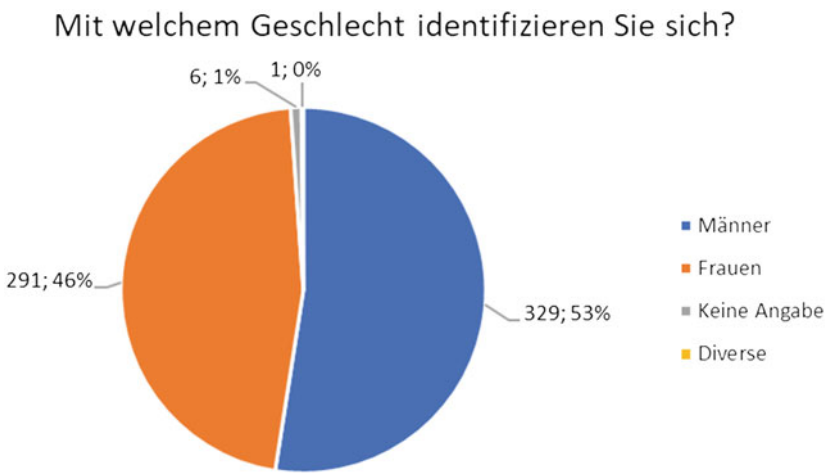

Abbildung 4.10 Angaben der Teilnehmenden zum Geschlecht (n = 627)

Abbildung 4.11 Angaben der Teilnehmenden zur Erstsprache (n = 627)

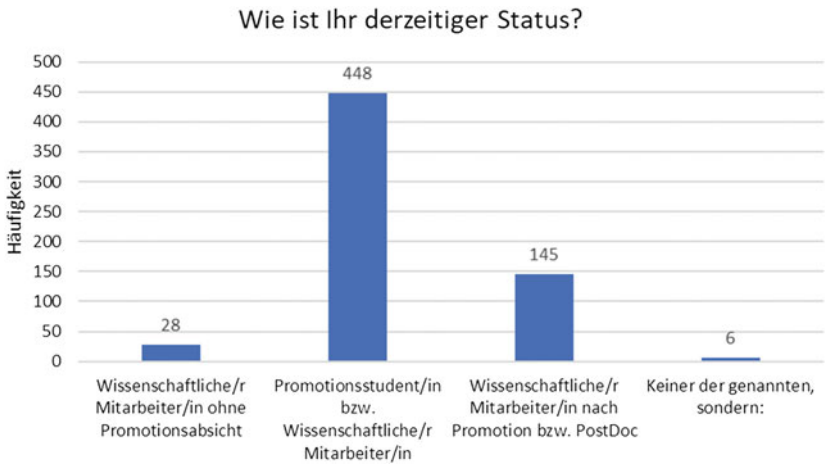

Abbildung 4.12 Angaben der Teilnehmenden zum Status (n = 627)

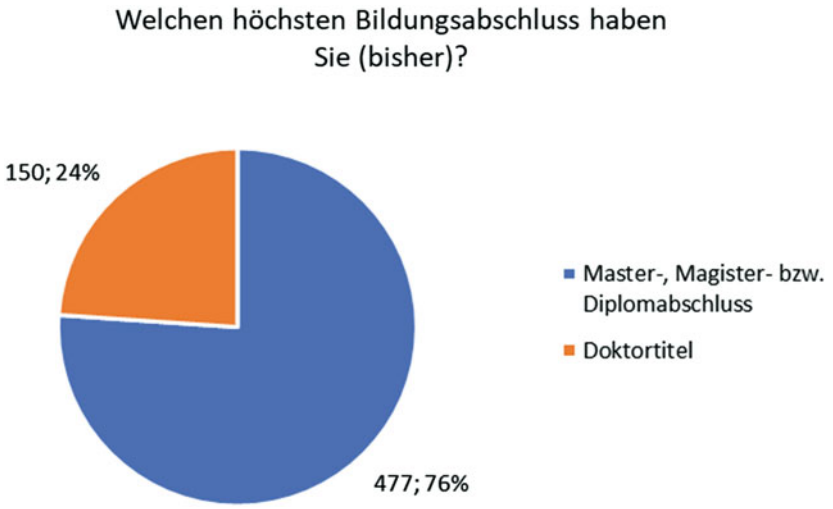

Abbildung 4.13 Angaben der Teilnehmenden zum Bildungsabschluss (n = 627)

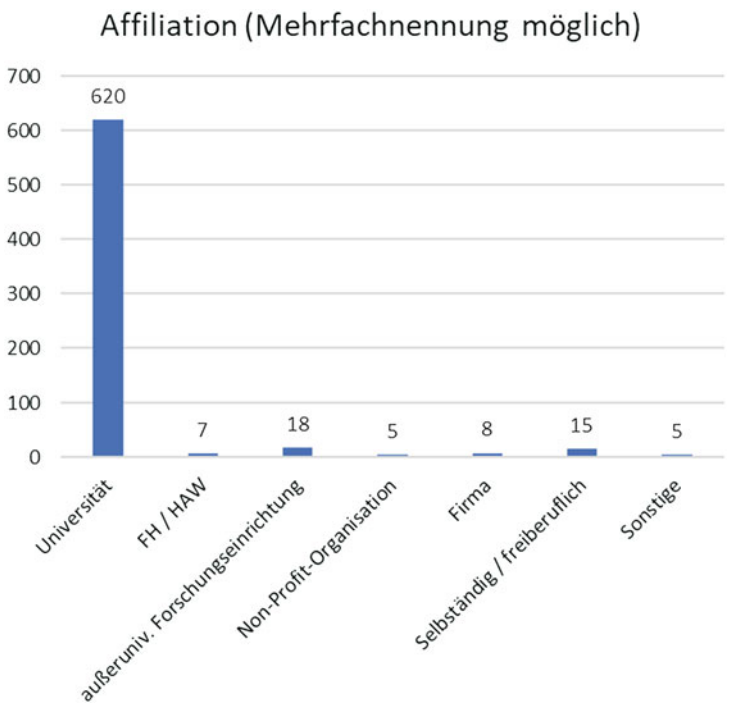

Abbildung 4.14 Angaben der Teilnehmenden zur Affiliation (n = 678)

4.4.2 Haupteffekte

Tabelle 4.11[45] enthält die Testergebnisse für die Haupteffekte und Interaktionseffekte, die alle jeweils einen statistisch signifikanten Wert von $p < 0{,}001$ aufweisen. Das heißt, es gibt mindestens einen statistisch signifikanten Unterschied auf mindestens zwei Stufen der jeweiligen unabhängigen Variablen. Demnach lassen sich zwar Effekte feststellen, für die Richtung (positiv oder negativ) und Stärke eines Effekts müssen die Differenzen zwischen den Mittelwerten (paarweise Vergleiche) für alle Stufen der UVn jeweils näher betrachtet werden.

[45] In dieser und allen folgenden Tabellen in diesem Abschnitt sind statistisch signifikante Unterschiede bei den Differenzen der Mittelwerte für eine bessere Übersichtlichkeit zusätzlich mit einem grünen Zellhintergrund versehen; nicht statistisch signifikante Werte, also $p > 0{,}05$, sind mit einem roten Zellhintergrund gekennzeichnet.

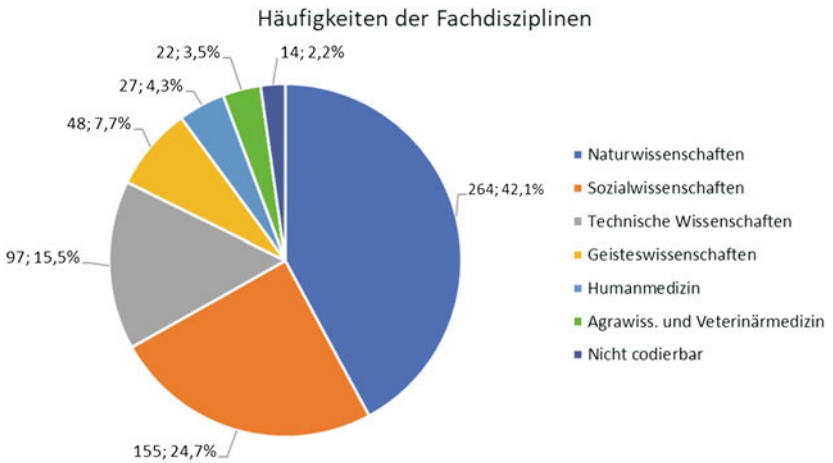

Abbildung 4.15 Anteile der in der Stichprobe vertretenen Fachdisziplinen (n = 627)

Im Folgenden werden zunächst die Wirkungen der einzelnen unabhängigen Variablen (Haupteffekte) untersucht, bevor in dem anschließenden Abschnitt 4.4.3 ihre Wechselwirkungen (Interaktionen) näher betrachtet werden.

Tabelle 4.11 Tests auf Haupt- und Interaktionseffekte, Typ III

Quelle	Zähler-Freiheitsgrade	Nenner-Freiheitsgrade	F-Wert	p-Wert
UV1	2	11154,09	98,92	< 0,001
UV2	2	11246,00	41,80	< 0,001
UV3	2	11274,65	69,18	< 0,001
UV1 * UV2	4	7885,62	214,15	< 0,001
UV1 * UV3	4	7511,41	39,87	< 0,001
UV2 * UV3	4	8586,08	7,73	< 0,001
UV1 * UV2 * UV3	8	5761,65	70,38	< 0,001

Anmerkung: UV 1 – Anzahl Downloads; UV 2 – Anzahl Zitationen Werk; UV 3 – Anzahl Zitationen Autor

4.4.2.1 Die Wirkung von UV 1 – Anzahl Downloads

Tabelle 4.12 zeigt die Mittelwerte der Bewertungen für UV 1 – Anzahl Downloads, die in Abbildung 4.16 grafisch dargestellt sind. Die Ergebnisse der paarweisen Vergleiche der Stufen für UV 1 untereinander sind in Tabelle 4.13 enthalten. An diesen Werten lässt sich ablesen, wo genau sich die statistisch signifikanten Unterschiede des Haupteffekts zeigen, die Differenz der Mittelwerte (I-J) zeigt jeweils das Ausmaß der Unterschiede an: Die durchschnittliche Punktzahl der Bewertungen fällt für Stufe 3 statistisch signifikant höher aus im Vergleich mit Stufe 1 (50,35 vs. 56,85) und Stufe 2 (51,26 vs. 56,85); im Vergleich zwischen Stufe 1 mit 2 gibt es keine signifikanten Unterschiede ($p = 0{,}189$). Somit besteht entgegen der in Hypothese 1 formulierten Vermutung ein negativer Effekt. Es ist nicht der Fall, dass bei Nichtanzeige (Stufe 3) die Punktzahl der Relevanzbewertung im Durchschnitt kleiner ist als bei einer geringen (Stufe 1) oder hohen (Stufe 2) Anzahl Downloads. Jedoch sind die Differenzen in den Mittelwerten von < 7 inhaltlich als relativ gering zu erachten. Vor diesem Hintergrund stellt sich die Frage, ab wann Unterschiede in den Bewertungen unabhängig von statistischer Signifikanz überhaupt auf einer inhaltlichen bzw. theoretischen Ebene als groß genug angesehen werden und von Bedeutung sein können. Auf diese Problematik wird im Zusammenhang mit den Interaktionseffekten in Abschnitt 4.4.3 genauer eingegangen.

Tabelle 4.12 Mittelwerte für UV 1 – Schätzungen

Stufen der Anzahl an Downloads (UV 1)	Mittelwert	Std.-Fehler
gering	50,35	0,35
hoch	51,26	0,35
keine Angabe	56,85	0,35

Tabelle 4.13 Mittelwerte für UV 1 – Paarweise Vergleiche

Stufen der Anzahl an Downloads (UV 1)		Differenz der	Std.-	
(I)	(J)	Mittelwerte (I-J)	Fehler	p-Wert
gering	hoch	−0,91	0,50	0,189
	keine Angabe	−6,51	0,50	< 0,001
Hoch	keine Angabe	−5,59	0,50	< 0,001

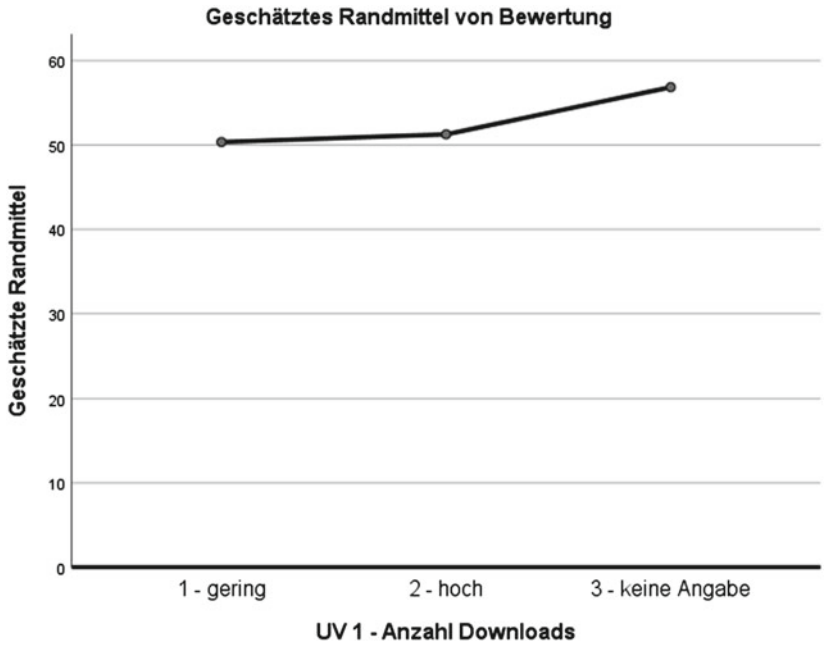

Abbildung 4.16 Mittelwerte der Bewertungen auf den Stufen von UV 1

4.4.2.2 Die Wirkung von UV 2 – Zitationen Werk

Die Mittelwerte der Bewertungen für UV 2 – Zitationen Werk zeigen Tabelle 4.14 und Abbildung 4.17. Auch hier liegen diese Werte sehr nah beieinander und zeigen entgegen der Vermutung, dass die Punktzahl der Bewertung bei einer hohen

Anzahl von Zitationen mit 52,18 im Durchschnitt zwar größer ist als bei einer geringen Anzahl (50,95), aber nicht größer als bei Nichtanzeige der Zitationszahl (55,34). Stattdessen sind die durchschnittlichen Bewertungen auf Stufe 3 – keine Angabe auch für UV 2 am höchsten. Wie in Tabelle 4.15 ersichtlich, sind die geringen Differenzen der Mittelwerte statistisch signifikant bei $p < 0{,}001$, im Vergleich der Stufen gering und hoch sind sie statistisch signifikant bei $p = 0{,}042$.

Tabelle 4.14 Mittelwerte für UV 2 – Schätzungen

Stufen der Zitationszahl Werk (UV 2)	Mittelwert	Std.-Fehler
gering	50,95	0,35
hoch	52,18	0,36
keine Angabe	55,34	0,35

Tabelle 4.15 Mittelwerte für UV 2 – Paarweise Vergleiche

Stufen der Zitationszahl Werk (UV 2)		Differenz der	Std.-	
(I)	(J)	Mittelwerte (I-J)	Fehler	p-Wert
gering	hoch	−1,23	0,50	0,042
	keine Angabe	−4,40	0,50	< 0,001
hoch	keine Angabe	−3,17	0,50	< 0,001

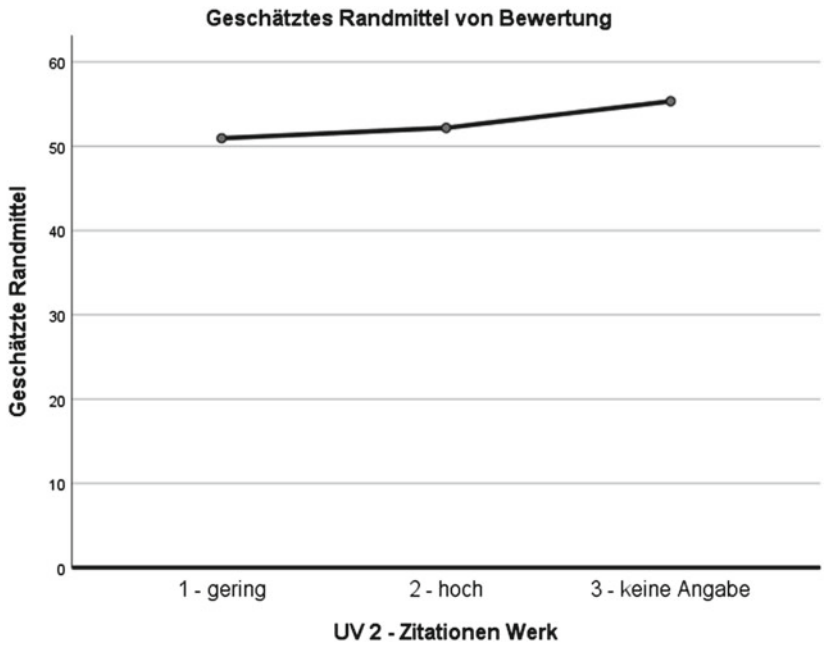

Abbildung 4.17 Mittelwerte der Bewertungen auf den Stufen von UV 2

4.4.2.3 Die Wirkung von UV 3 – Zitationen Autor

Während die Durchschnittsbewertungen für UV 1 – Anzahl Downloads und UV 2 – Zitationszahl Werk in Abbildung 4.17 und Abbildung 4.17 dasselbe Muster zeigen, weicht es für UV 3 – Zitationszahl Autor in Abbildung 4.18 aufgrund der Werte in Stufe 3 – keine Angabe leicht davon ab: Die Punktzahl der Bewertung ist bei einer geringen Anzahl Zitationen des Autors im Durchschnitt kleiner (49,40) als bei einer hohen Anzahl (54,63) und bei einer hohen Anzahl etwas größer als bei Nichtanzeige; allerdings ist sie bei Nichtanzeige entgegen der Erwartung mit 54,43 höher als bei einer geringen Anzahl (vgl. Tabelle 4.16). Die Mittelwertdifferenzen (vgl. Tabelle 4.17) sind auch im Vergleich der Stufen von UV 3 gering, aber für Stufe 1 – gering mit Stufe 2 – hoch und Stufe 3 – keine Angabe statistisch signifikant ($p < 0{,}001$). Die äußerst geringe Differenz der Mittelwerte von Stufe 2 und Stufe 3 weist hingegen keine statistische Signifikanz auf.

Tabelle 4.16 Mittelwerte für UV 3 – Schätzungen

Stufen Zitationszahl Autor (UV 3)	Mittelwert	Std.-Fehler
gering	49,40	0,36
hoch	54,63	0,35
keine Angabe	54,43	0,35

Tabelle 4.17 Mittelwerte für UV 3 – Paarweise Vergleiche

Stufen Zitationszahl Autor (UV 3) (I)	(J)	Differenz der Mittelwerte (I-J)	Std.-Fehler	p-Wert
gering	hoch	−5,23	0,50	< 0,001
	keine Angabe	−5,02	0,50	< 0,001
hoch	keine Angabe	0,21	0,50	0,966

Zusammengefasst lässt sich feststellen, dass die unabhängigen Variablen auf Stufe 3 – keine Angabe entgegen den Erwartungen im Durchschnitt höhere Relevanzbewertungen bewirken als auf den Stufen 1 – gering und 2 – hoch. Die Annahmen über die Haupteffekte lassen sich nicht in Gänze bestätigen; allerdings können keine allgemeingültigen Aussagen zu den Wirkungen der unabhängigen Variablen getroffen werden, weil ebenfalls statistisch signifikante Interaktionen vorliegen, die näher zu untersuchen sind.

4.4.3 Interaktionseffekte

Für die in diesem Experiment untersuchten drei unabhängigen Variablen (UVn) konnten sowohl Interaktionseffekte der 1. Ordnung (2-fach-Interaktionen) als auch Interaktionseffekte der 2. Ordnung (3-fach-Interaktionen) nachgewiesen werden.

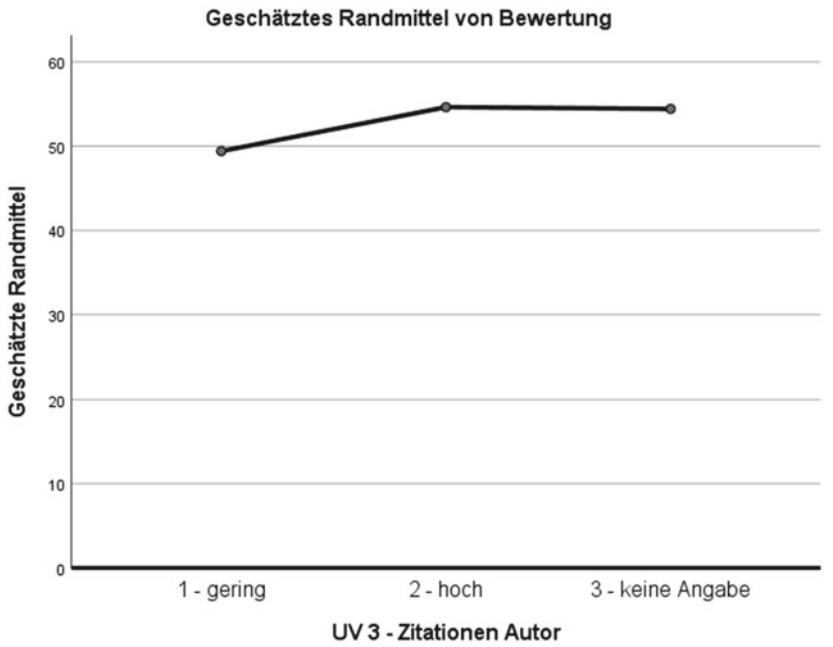

Abbildung 4.18 Mittelwerte der Bewertungen auf den Stufen von UV 3

Das Vorhandensein von Interaktionen lässt sich (üblicherweise bei unabhängigen Variablen mit zwei Kategorien) anhand grafischer Darstellungen in Form von Liniendiagrammen gut ablesen, obwohl die Linien in den Diagrammen suggerieren, dass es Messwerte zwischen den Stufen der UV gibt, die gar nicht erhoben wurden (Sedlmeier & Renkewitz, 2018, S. 173, 174). Würden die Linien der Faktoren parallel verlaufen, lägen keine Wechselwirkungen, sondern eine Nullinteraktion vor, d. h. die Haupteffekte eines Faktors wären auf allen Stufen eines anderen Faktors gleich groß. Dies ist weder für die hier vorliegenden 2-fach-Interaktionen noch für die 3-fach-Interaktionen bis auf wenige Ausnahmen der Fall. Stattdessen wird bereits anhand der beiden Liniendiagramme für die Darstellung der Wechselwirkungen zwischen zwei UVn ein komplexes Ergebnismuster sichtbar: Jede UV besitzt drei Kategorien (gering – hoch – keine Angabe), deren Linien häufig zwischen den Kategorien *gering – hoch* und den Kategorien *hoch – keine Angabe* gegensätzliche Richtungen in unterschiedlich starken Anstiegen aufzeigen. Für eine bessere Unterscheidung der Linienverläufe werden in allen

nachfolgenden Diagrammen die Linien auf den UV-Stufen *gering* und *hoch* von denen auf der Stufe *keine Angabe* getrennt beschrieben. Diese separate Betrachtung der Kategorie *keine Angabe* soll jeweils mithilfe der gestrichelten vertikalen Linie bei Stufe 2 in den Diagrammen vereinfacht werden. Konkret bedeutet dies, dass zunächst der Vergleich der Mittelwerte auf den Stufen *gering* und *hoch* erfolgt und im Anschluss der Vergleich auf den Stufen *hoch* und *keine Angabe* sowie *gering* und *keine Angabe*, wodurch ein Vergleich der Mittelwerte in jeder Stufenkombination hergestellt werden kann.

Für die Beschreibung der Interaktionseffekte sei an dieser Stelle erneut auf die Frage hingewiesen, ab wann ein Bewertungsunterschied zwischen den Stufen einer UV im Vergleich zu einer anderen UV (2-fach-Interaktion) bzw. zu den anderen beiden UVn (3-fach-Interaktion) als groß genug gilt, unabhängig davon, ob der Unterschied statistisch signifikant ist oder nicht. Im vorangegangenen Abschnitt 4.4.2 wurden die Haupteffekte anhand der Mittelwerte und der paarweisen Vergleiche beschrieben. Es wurde festgestellt, dass die Unterschiede in den Bewertungen auf den Stufen der einzelnen UVn untereinander überaus gering waren – die größte Differenz mit −6,51 weist UV 1 bei dem Vergleich der Messwerte in Stufe 1 – gering mit Stufe 3 – keine Angabe auf (vgl.

Tabelle 4.13). Sehr geringe Unterschiede von zwei oder drei Punkten lassen sich vermutlich weniger auf inhaltlich begründete, intendierte Bewertungen zurückführen, sondern mit hoher Wahrscheinlichkeit auf die Handhabung des Schiebereglers mit der Maus. Ob auch Unterschiede von fünf oder sieben Punkten diesem Umstand geschuldet sind, kann nur spekuliert werden.

Obwohl solche geringen Unterschiede teilweise statistisch signifikant sind (beispielsweise die Differenz von −1,23 bei UV 2 in den Stufen gering und hoch mit $p = 0{,}042$, vgl. Tabelle 4.15), scheint es naheliegend, diese als inhaltlich unbedeutend zu beurteilen. Vor diesem Hintergrund werden derart geringe Unterschiede inhaltlich gleichgesetzt und in den nachfolgenden Ausführungen vernachlässigt.[46] Stattdessen werden Differenzen in den Bewertungen von mindestens 10 und 20 Punkten vornehmlich betrachtet. Diese Schwellen bei den Differenzen lassen sich mit ihrem Verhältnis zur Anzahl der Abstufungen der verwendeten Schiebereglerskala argumentieren: Unterschiede von 10 Punkten nehmen ein Zehntel der Gesamtskala ein, Unterschiede von 20 Punkten sind ein Fünftel der Gesamtskala; im Vergleich mit einer 5-Punkte-Skala entsprechen

[46] Zu beachten ist, dass sich diese inhaltlich bedeutsamen Differenzwerte auf die paarweisen Vergleiche beschränken und in den Liniendiagrammen alle Mittelwerte mit einbezogen sind. Das bedeutet, dass in den Diagrammen weder zwischen statistisch noch inhaltlich bedeutsamen Werten unterschieden werden kann.

20 Punkte einem Punkt. Differenzen dieser Größe bieten daher eine sinnvollere Grundlage für die Interpretation der Ergebnisse.

In diesem Abschnitt wird auf eine ausführliche Darstellung der 2-fach-Interaktionen[47] verzichtet, stattdessen wird der Fokus auf die 3-fach-Interaktionen gelegt. Wie bereits bei den Haupteffekten festgestellt, entsprechen die Wechselwirkungen zwischen UV 1 und UV 2, UV 2 und UV 3 sowie UV 1 und UV 3 nicht den Erwartungen. Vor dem Hintergrund, dass die Nichtanzeige von Download- oder Zitationszahlen im Durchschnitt zu höheren Relevanzbewertungen führt als die Anzeige geringer oder höherer Zahlen, wird zusätzlich ein komplexes Ergebnismuster sichtbar, welches keine generalisierenden Aussagen über die Wirkung einer einzelnen UV auf den Stufen einer zweiten UV zulässt. Es ist zu prüfen, ob sich hinsichtlich der 3-fach-Interaktionen das komplexe Ergebnismuster fortsetzt oder sich ein klareres Bild abzeichnet, auf dessen Basis Aussagen über die Wirkungen der einzelnen unabhängigen Variablen auf den Stufen aller weiteren unabhängigen Variablen getroffen werden können.

In den nachfolgenden Abschnitten 4.4.3.1, 4.4.3.2 und 4.4.3.3 werden die paarweisen Vergleiche der Wechselwirkungen aller drei UVn und insbesondere die Differenzwerte, die mehr als 10 bzw. 20 Bewertungspunkte aufweisen, berichtet. Da sich die Schätzungen der Mittelwerte für die 3-fach-Interaktion lediglich in der Reihenfolge der Anzeige unterscheiden, wird auf eine redundante Darstellung der Mittelwerte pro Wechselwirkung verzichtet. Am Ende jedes Abschnitts werden die Interaktionen mithilfe der Liniendiagramme erläutert.

Tabelle 4.18 zeigt die Mittelwerte der Bewertungen in allen 27 experimentellen Bedingungen, die einen Gesamtmittelwert von 52,82 bilden, wobei der Median bei 54,36 liegt. Zusätzlich beinhaltet sie die Nummer der jeweiligen experimentellen Bedingung (Merkmalskombination), durch die die Kombination aller Stufen der drei unabhängigen Variablen eine eindeutige Bezeichnung aufweist. Den höchsten Wert bietet die Merkmalskombination S322 mit 67,29 in der Kombination UV 1 – keine Angabe, UV 2 – hoch und UV 3 – hoch, den kleinsten Mittelwert mit 29,87 weist S221 auf in der Kombination UV 1 – hoch, UV 2 – hoch, UV 3 – gering. Die Kombination gering – gering – gering (S111) zeigt mit 41,71 einen wesentlich kleineren Mittelwert als die Kombination keine Angabe – keine Angabe – keine Angabe (S333) mit einem Wert von 51,30; die Kombination hoch – hoch – hoch (S222) hat einen Mittelwert von 50,08, welcher etwas geringer ausfällt als die durchschnittliche Bewertung in S333. Derartige

[47] Die Ergebnisse der 2-fach-Interaktionen zwischen den UV 1 und UV 2, UV 1 und UV 3 sowie UV 2 und UV 3 sowie deren Erläuterungen sind in Anhang 4.1 im elektronischen Zusatzmaterial enthalten.

Abweichungen von den Hypothesen zeigten sich bereits bei den Betrachtungen der Einzelwirkungen (Haupteffekte) und den Wechselwirkungen der 1. Ordnung.

Die Verteilung der Mittelwerte über alle 27 Teilstichproben mit Gesamtmittelwert und Median ist in Abbildung 4.19 grafisch dargestellt. Die Zuordnung einer Bedingung zu einer Aufgabe ist in Klammern angegeben. Gut erkennbar ist die Lage der Mittelwerte von S111, S222 und S333, die alle sowohl unter dem Gesamtmittelwert (52,82) als auch unter dem Median (54,36) liegen.

Bevor die jeweiligen Wirkungen aller drei unabhängigen Variablen auf den Stufen der jeweils anderen zwei UVn beschrieben werden, bietet Tabelle 4.19 eine Übersicht über die quantitative Verteilung der insgesamt 36 inhaltlich bedeutsamen Differenzwerte. Diese zeigt bereits die weniger starke Wirkung von UV 3 – Zitationszahl Autor im Vergleich zu den Wirkungen von UV 1 – Anzahl Downloads und UV 2 – Zitationszahl Werk: So weisen letztere insgesamt eine höhere Anzahl bedeutender Unterschiede auf, von denen sogar jeweils ein Wert bei ≥ 30 liegt.

4.4.3.1 Die Wirkung von UV 1 – Anzahl Downloads auf den Stufen von UV 2 – Zitationszahl Werk und UV 3 – Zitationszahl Autor

Das Ausmaß der Unterschiede der Bewertungen von UV 1 – Anzahl Downloads auf den Stufen von UV 2 – Zitationszahl Werk zeigen die paarweisen Vergleiche in Tabelle 4.20[48]. Im Gegensatz zu den paarweisen Vergleichen für die Interaktionseffekte 1. Ordnung werden nun deren Wechselwirkungen zusätzlich auf jeder Stufe von UV 3 – Zitationszahl Autor betrachtet. Die Hinzunahme der dritten UV führt dazu, dass die Tabellen jeweils 27 Differenzwerte beinhalten, während es für die 2-fach-Interaktionen im vorangegangenen Abschnitt nur jeweils 9 Werte waren. Obwohl diese nun vollständige Abdeckung von Differenzwerten zu einer ebenso vollständigen Betrachtung der einzelnen Differenzen verleiten mag, werden die inhaltlich bedeutsamen Differenzwerte, d. h. Werte größer 10 bzw. 20, in den Fokus genommen.

[48] In dieser und allen nachfolgenden Tabellen dieses Abschnitts**.** sind Differenzwerte größer gleich 10 bzw. größer gleich 20 mit einem farbigen Hintergrund in Hellblau bzw. Lila gekennzeichnet.

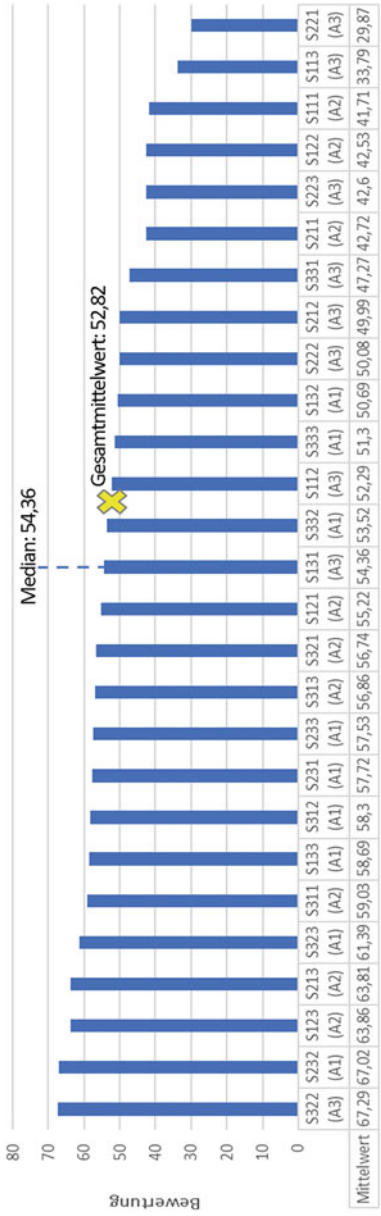

Abbildung 4.19 Diagramm der Mittelwerte aus den Bewertungen aller 27 Bedingungen (Teilstichproben)

Tabelle 4.18 Mittelwerte für UV 1 * UV 2 * UV 3 – Schätzungen

Nr. Bedingung	Stufen der Anzahl Downloads (UV 1)	Stufen der Zitationszahl Werk (UV 2)	Stufen der Zitationszahl Autor (UV 3)	Mittelwert	Std.-Fehler
S111	gering	gering	gering	41,71	1,06
S112			hoch	52,29	1,12
S113			keine Angabe	33,79	0,94
S121		hoch	gering	55,22	1,07
S122			hoch	42,53	1,11
S123			keine Angabe	63,86	1,06
S131		keine Angabe	gering	54,36	1,13
S132			hoch	50,69	1,03
S133			keine Angabe	58,69	0,99
S211	hoch	gering	gering	42,72	1,07
S212			hoch	49,99	1,08
S213			keine Angabe	63,81	1,06
S221		hoch	gering	29,87	1,05
S222			hoch	50,08	1,09
S223			keine Angabe	42,60	1,06
S231		keine Angabe	gering	57,72	1,03
S232			hoch	67,02	1,01
S233			keine Angabe	57,53	1,09
S311	keine Angabe	gering	gering	59,03	1,00
S312			hoch	58,30	1,04
S313			keine Angabe	56,86	1,04
S321		hoch	gering	56,74	1,17
S322			hoch	67,29	1,01
S323			keine Angabe	61,39	1,10
S331		keine Angabe	gering	47,27	1,09
S332			hoch	53,52	1,04
S333			keine Angabe	51,30	1,08

Tabelle 4.19 Inhaltlich bedeutsame Differenzwerte der 3-fach-Interaktionen

	UV 1 – Anzahl Downloads (Tabelle 4.20)	UV 2 – Zitationszahl Werk (Tabelle 4.21)	UV 3 – Zitationszahl Autor (Tabelle 4.22)	Gesamt
Differenzwerte ≥ 10	7	10	6	23
Differenzwerte ≥ 20 (davon ≥ 30)	6 (1)	4 (1)	3 (0)	13
Gesamt	13	14	9	36

Tabelle 4.20 zeigt, dass 7 der 27 paarweisen Vergleiche nicht statistisch signifikant sind, 20 hingegen sind signifikant bei $p < 0{,}001$. Die nicht statistisch signifikanten Werte sind jedoch ohnehin vernachlässigbar gering, während unter den statistisch signifikanten Werten 7 Differenzen größer als 10 sind, größer als 20 sind 6 Differenzwerte und somit sind 13 Werte auch von inhaltlicher Bedeutung. Diese Werte verteilen sich über die Stufen von UV 3 – Zitationszahl Autor und UV 2 – Zitationszahl Werk folgendermaßen:

- Bei einer *geringen Anzahl* an Autoren-Zitationen (UV 3 – Stufe 1) sind 3 Werte > 10 und 2 Werte > 20, wobei letztere jeweils bei einer hohen Anzahl Werks-Zitationen (UV 2) auftreten: Die Relevanzbewertung fällt hier entgegen den Erwartungen um durchschnittlich 25,35 Bewertungspunkte höher aus, wenn die Anzahl der Downloads (UV 1) gering ist im Vergleich zu einer hohen Anzahl Downloads und um durchschnittlich 26,87 Punkte höher, wenn die Anzahl der Downloads nicht angezeigt wird im Vergleich zur Anzeige einer hohen Anzahl Downloads. Bei einer geringen Anzahl Werks-Zitationen (UV 2) fällt die Relevanzbewertung im Durchschnitt um 17,32 Punkte größer aus, wenn die Anzahl der Downloads nicht angezeigt wird im Vergleich zu einer geringen Anzahl Downloads und ebenso um 16,31 Punkte größer im Vergleich zu einer hohen Anzahl Downloads. Wird die Anzahl der Zitationen eines Werks nicht angezeigt, ist die Relevanzbewertung im Mittel um 10,46 Punkte größer, wenn die Anzahl der Downloads hoch ist im Vergleich zur Nichtanzeige.
- Bei einer *hohen Anzahl* an Autoren-Zitationen (UV 3 – Stufe 2) sind 3 Werte > 10 und 1 Wert > 20, letztere tritt ebenfalls bei einer hohen Anzahl Werks-Zitationen (UV 2) auf: Die Bewertung fällt um durchschnittlich 24,76 Punkte höher aus, wenn die Anzahl der Downloads (UV 1) nicht angezeigt wird im Vergleich zur Anzeige einer geringen Anzahl Downloads. Zudem ist

die Bewertungspunktzahl im Durchschnitt 17,21 größer bei Nichtanzeige der Anzahl der Downloads im Vergleich zur Anzeige einer hohen Anzahl Downloads. Ist die Anzahl der Werks-Zitationen (UV 2) nicht angegeben, ist bei einer hohen Anzahl Downloads die Bewertung im Durchschnitt 16,34 Punkte höher im Vergleich zu einer geringen Anzahl Downloads und 13, 51 Punkte größer als bei Nichtanzeige der Downloadzahl.

- Bei *Nichtanzeige* der Autoren-Zitationen (UV 3 – Stufe 3) sind 4 Werte inhaltlich relevant, wobei 1 Wert > 10 und 3 Werte > 20 sind. Die größte durchschnittliche Differenz liegt bei einer geringen Anzahl Werks-Zitationen: Die Bewertung fällt im Mittel um 30,02 Punkte höher aus bei einer hohen Anzahl Downloads und um 23,07 Punkte höher bei Nichtanzeige im Vergleich zu einer geringen Anzahl Downloads. Bei einer hohen Anzahl Werks-Zitationen fällt die Punktzahl der Relevanzbewertung im Durchschnitt um 21,26 Punkte kleiner aus, wenn die Anzahl der Downloads hoch ist im Vergleich zu einer geringen Anzahl Downloads; die Bewertung ist durchschnittlich 18,79 Punkte höher, wenn die Anzahl der Downloads nicht angegeben ist im Vergleich zur Anzeige einer hohen Anzahl Downloads. Bei Nichtanzeige der Werks-Zitationen liegen die Differenzwerte unter zehn Bewertungspunkten und können daher vernachlässigt werden.

Hervorzuheben ist hier der Differenzwert von −30,02 bei einem Vergleich der Kategorien *gering* und *hoch* bei UV 1 – Anzahl Downloads auf den Stufen 1 – gering von UV 2 – Zitationszahl Werk und 3 – keine Angabe von UV 3 – Zitationszahl Autor. Dieser Wert ist konform mit der Annahme, dass eine hohe Anzahl Downloads (UV 1) zu einer höheren Relevanzbewertung führt als eine geringe Anzahl Downloads. Ähnliches ist der Fall in der Kombination *gering* und *hoch* bei UV 1 auf der Stufe 3 – keine Angabe von UV 2 – Zitationszahl Werk und auf der Stufe 2 – hoch von UV 3 – Zitationszahl Autor mit dem Wert −6,34. Dagegen finden sich bei zwei Kombinationen von UV 1 – Anzahl Downloads Differenzwerte, die eine höhere Durchschnittsbewertung für die Stufe 1 – gering im Vergleich mit der Stufe 2 – hoch aufweisen: Dies betrifft den Vergleich von *gering* und *hoch* bei Stufe 2 – hoch von UV 2 und bei Stufe 1 – gering von UV 3 (25,35) sowie bei Stufe 2 – hoch von UV 2 – Zitationszahl Werk und auf Stufe 3 – keine Angabe bei UV 3 – Zitationszahl Autor (21,26). Zudem fällt auf, dass UV 3 auf Stufe 3 – keine Angabe die häufigsten Differenzwerte > 20 bei Stufe 1 – gering von UV 2 und Stufe 2 – hoch bei UV 2 hervorruft.

Die Liniendiagramme zur Darstellung der 3-fach-Interaktionen zeigen die Wirkung der jeweiligen UV (auf der y-Achse) auf den Stufen einer zweiten UV (auf der x-Achse) für jede Stufe der dritten UV in einem separaten Diagramm.

In Abbildung 4.20 sind die durchschnittlichen Bewertungen von UV 1 – Anzahl Downloads jeweils auf den Stufen von UV 2 – Zitationszahl Werk und UV 3 – Zitationszahl Autor dargestellt. Im ersten Diagramm der oberen Reihe verlaufen bei einer geringen Anzahl Autoren-Zitationen die Linien von UV 1 auf der Stufe 1 – gering im Vergleich mit Stufe 2 – hoch von UV 2 erneut gegensätzlich und entgegen der Erwartung: Die Linie *gering* weist einen starken positiven Anstieg auf, die Linie *hoch* zeigt eine eindeutige negative Richtung, während die Linie der Kategorie *keine Angabe* leicht negativ verläuft. Auf den Stufen 2 – hoch und 3 – keine Angabe von UV 2 (rechts der Trennlinie) weist die Linie *hoch* hingegen einen starken positiven Anstieg auf, während die Linien *gering* leicht negativ und *keine Angabe* negativ verlaufen. Bei dem Vergleich von Stufe 1 – gering und 3 – keine Angabe liegen die Kategorien *gering* und *hoch* mit einem ähnlich positiven Anstieg zueinander, die Kategorie *keine Angabe* hingegen verläuft negativ.

Im ersten Diagramm der unteren Reihe zeigt die Linie *keine Angabe* bei einer geringen Anzahl Werks-Zitationen einen ähnlich negativen Verlauf auf den Stufen von UV 3 – Zitationszahl Autor. Die Linien *gering* und *hoch* verlaufen nahezu gleich positiv auf der Stufe 1 – gering und 2 – hoch von UV 3, wohingegen die Linie *hoch* im Vergleich der Stufen 2 – hoch und 3 – keine Angabe stark ansteigt, die Linie *gering* stark abfällt. Von den Stufen 1 und 3 her betrachtet verhalten sich die Kategorien *gering* leicht negativ und *hoch* stark positiv.

Bei einer hohen Anzahl an Autoren-Zitationen (im zweiten Diagramm der oberen Reihe) verläuft die Linie *hoch* auf den Stufen 1 – gering und 2 – hoch von UV 2 – Zitationszahl Werk weder positiv noch negativ, während die Linie *gering* leicht negativ und die Linie *keine Angabe* positiv verläuft. Bei dem Vergleich von Stufe 2 – hoch und 3 – keine Angabe rechts von der Trennlinie zeigen die Linien *gering* und *hoch* entgegen der Erwartung einen positiven Anstieg, die Linie *keine Angabe* verläuft hingegen erwartungsgemäß negativ. Die Unterschiede der Kategorien *gering* und *keine Angabe* bei dem Vergleich der Stufen 1 – gering und 3 – keine Angabe von UV 2 sind nur minimal negativ, die Kategorie *hoch* verhält sich dagegen deutlich positiv.

Tabelle 4.20 Mittelwerte für UV 3 * UV 2 * UV 1 – Paarweise Vergleiche

Stufen der Zitationszahl Autor (UV 3)	Stufen der Zitationszahl Werk (UV 2)	Stufen der Anzahl Downloads (UV 1) (I)	(J)	Differenz der Mittelwerte (I-J)	Std.-Fehler	p-Wert
gering	gering	gering	hoch	−1,01	1,50	0,877
			keine Angabe	−17,32	1,46	< 0,001
		hoch	keine Angabe	−16,31	1,46	< 0,001
	hoch	gering	hoch	25,35	1,50	< 0,001
			keine Angabe	−1,52	1,58	0,707
		hoch	keine Angabe	−26,87	1,58	< 0,001
	keine Angabe	gering	hoch	−3,37	1,53	0,080
			keine Angabe	7,09	1,56	< 0,001
		hoch	keine Angabe	10,46	1,50	< 0,001
hoch	gering	gering	hoch	2,30	1,55	0,360
			keine Angabe	−6,00	1,53	< 0,001
		hoch	keine Angabe	−8,31	1,50	< 0,001
	hoch	gering	hoch	−7,55	1,56	< 0,001
			keine Angabe	−24,76	1,50	< 0,001
		hoch	keine Angabe	−17,21	1,49	< 0,001
	keine Angabe	gering	hoch	−16,34	1,44	< 0,001
			keine Angabe	−2,83	1,46	0,150
		hoch	keine Angabe	13,51	1,44	< 0,001
keine Angabe	gering	gering	hoch	−30,02	1,42	< 0,001
			keine Angabe	−23,07	1,41	< 0,001
		hoch	keine Angabe	6,95	1,49	< 0,001
	hoch	gering	hoch	21,26	1,50	< 0,001
			keine Angabe	2,47	1,53	0,285
		hoch	keine Angabe	−18,79	1,52	< 0,001
	keine Angabe	gering	hoch	1,16	1,47	0,817
			keine Angabe	7,39	1,47	< 0,001
		hoch	keine Angabe	6,23	1,54	< 0,001

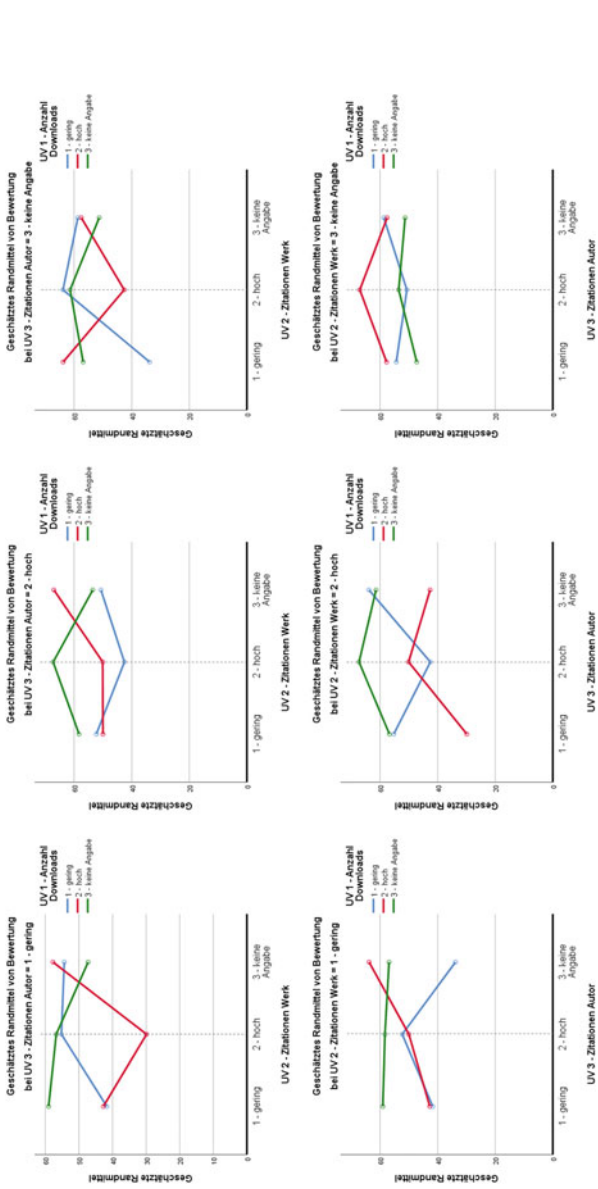

Abbildung 4.20 Mittelwerte der Bewertungen von UV 1 auf den Stufen von UV 2 bei UV 3 (obere Reihe) und auf den Stufen von UV 3 bei UV 2 (untere Reihe)

Im zweiten Diagramm der unteren Reihe verlaufen die Linien *hoch* und *keine Angabe* auf den Stufen 1 – gering und 2 – hoch von UV 3 – Zitationszahl Autor bei einer hohen Anzahl Werks-Zitationen (UV 2) nahezu parallel mit einem positiven Anstieg, die Linie *gering* verläuft negativ. Rechts von der Trennlinie, auf den Stufen 2 – hoch und 3 – keine Angabe von UV 3, setzt sich der parallele Verlauf der Linien *hoch* und *keine Angabe* fort, allerdings in negativer Richtung, während die Linie *gering* stark ansteigt. Der Vergleich der Kategorien auf den Stufen 1 – gering und 3 – keine Angabe zeigt, dass alle drei Linien nahezu parallel verlaufen.

Im dritten Diagramm der oberen Reihe, bei Nichtanzeige der Autoren-Zitationen, verläuft die Linie *keine Angabe* auf den Stufen 1 – gering und 2 – hoch von UV 2 – Zitationszahl Werk leicht positiv, die Linie *gering* dagegen stark positiv und die Linie *hoch* stark negativ. Rechts von der Trennlinie, auf den Stufen 2 – hoch und 3 – keine Angabe von UV 2, verlaufen die Linien *gering* und *keine Angabe* leicht negativ, während die Linie *hoch* einen starken Anstieg aufweist. Bei dem Vergleich der Kategorien auf den Stufen 1 – gering und 3 – keine Angabe zeigt sich, dass die Linien *hoch* und *keine Angabe* nahezu parallel, also ohne Interaktion, leicht negativ verlaufen, während die Linie *gering* stark positiv verläuft.

Ebenfalls nahezu parallel verlaufen im dritten Diagramm der unteren Reihe die Linien *hoch* und *keine Angabe* mit einem leicht positiven Anstieg auf den Stufen 1 – gering und 2 – hoch von UV 3 – Zitationszahl Autor, die Linie *gering* zeigt einen leicht negativen Verlauf. Auf den Stufen 2 – hoch und 3 – keine Angabe von UV 3 (rechts von der Trennlinie) sind die Linien *hoch* und *keine Angabe* leicht negativ gerichtet, während die Linie *gering* leicht positiv verläuft. Der Vergleich der Kategorien von UV 1 – Anzahl Downloads auf den Stufen 1 – gering und 3 – keine Angabe von UV 3 – Zitationszahl Autor zeigt erneut einen annähernd parallelen Verlauf für die Linien *gering* und *keine Angabe* bei einem leicht positiven Anstieg, die Linie *hoch* zeigt einen kaum merklichen Unterschied auf.

4.4.3.2 Die Wirkung von UV 2 – Zitationszahl Werk auf den Stufen von UV 1 – Anzahl Downloads und UV 3 – Zitationszahl Autor

In diesem Abschnitt werden die Wechselwirkungen von UV 2 – Zitationszahl Werk auf den Stufen von UV 1 – Anzahl Downloads und wiederum auf den einzelnen Stufen von UV 3 – Zitationszahl Autor betrachtet. Die Ergebnisse der paarweisen Vergleiche in Tabelle 4.21 zeigen, dass lediglich vier Differenzwerte nicht statistisch signifikant sind, während von den anderen Werten 19 statistisch

signifikant sind bei $p < 0{,}001$ und vier Werte bei $p < 0{,}05$. 10 Differenzwerte sind im Durchschnitt > 10, bei 4 Werten sind die Differenzen > 20. Diese 14 inhaltlich bedeutsamen Werte verteilen sich über die Stufen von UV 3 und UV 2 folgendermaßen:

- Bei einer *geringen Anzahl* an Autoren-Zitationen (UV 3 – Stufe 1) sind 5 Differenzwerte > 10, bei einem paarweisen Vergleich ist die Differenz größer als 20. Hier ist der höchste Differenzwert zu finden bei einer hohen Anzahl an Downloads: Die Relevanzbewertung fällt im Durchschnitt 27,85 Punkte höher aus bei Nichtanzeige der Werks-Zitationen im Vergleich zu einer hohen Anzahl; ebenfalls höher ist die Bewertung bei Nichtanzeige im Vergleich zur Anzeige einer geringen Anzahl Werks-Zitationen (im Durchschnitt 15,00 Punkte) während sie bei einer hohen Anzahl im Durchschnitt 12,85 Punkte kleiner ausfällt im Vergleich zu einer geringen Anzahl Werks-Zitationen. Bei einer geringen Anzahl Downloads treten beide Differenzwerte > 10 bei dem Vergleich mit einer geringen Anzahl Werks-Zitationen auf: Die Relevanzbewertung ist durchschnittlich 13,51 Punkte höher bei einer hohen Anzahl Werks-Zitationen und 12,65 Punkte höher, wenn die Anzahl Werks-Zitationen nicht angegeben ist. Ist hingegen die Anzahl der Downloads nicht angegeben, gibt es nur einen inhaltlich bedeutsamen Wert: Die Punktzahl der Relevanzbewertung ist im Mittel 11,76 höher, wenn die Anzahl der Werks-Zitationen gering ist im Vergleich zu deren Nichtanzeige.
- Bei einer *hohen Anzahl* an Autoren-Zitationen (UV 3 – Stufe 2) liegen drei Werte über der 10-Punkte-Schwelle. Der höchste Differenzwert ist zu finden bei einer hohen Anzahl an Downloads (UV 1 – Stufe 2): Die Bewertung ist im Mittel 17,03 Punkte kleiner bei einer geringen Anzahl Werks-Zitationen und ebenfalls kleiner um 16,95 Punkte bei einer hohen Anzahl im Vergleich zu der Nichtanzeige der Werks-Zitationen. Ist die Anzahl der Downloads nicht angegeben, ist die Punktzahl der Relevanzbewertung im Durchschnitt 13,77 Punkte größer, wenn die Anzahl der Werks-Zitationen hoch ist im Vergleich zur Nichtanzeige der Werks-Zitationen.
- Bei *Nichtanzeige* der Autoren-Zitationen (UV 3 – Stufe 3) sind für die Interaktionen der drei UVn auch in dieser Konstellation (analog zu den Differenzwerten in Tabelle 4.20) die meisten Differenzwerte > 20 festzustellen, von denen zwei Werte auf der Stufe 1 – gering von UV 1 – Anzahl Downloads liegen: Die Bewertung fällt um 30,07 Punkte im Durchschnitt größer aus, wenn die Anzahl der Werks-Zitationen hoch ist und um 24,99 Punkte größer bei deren Nichtanzeige im Vergleich zu einer geringen Anzahl Werks-Zitationen. Ein weiterer Differenzwert > 20 liegt bei einer hohen Anzahl an

Downloads (UV 1 – Stufe 2): Die Punktzahl der Bewertung fällt im Mittel 21,21 kleiner aus, wenn die Anzahl der Werks-Zitationen hoch ist im Vergleich zu einer geringen Anzahl an Werks-Zitationen. Ebenfalls bei einer hohen Anzahl Downloads ist die Bewertung um 14,93 Punkte höher, wenn die Anzahl der Werks-Zitationen nicht angegeben ist im Vergleich zu einer hohen Anzahl Werks-Zitationen. Bei Nichtanzeige der Anzahl an Downloads fällt die Bewertung im Durchschnitt 10,10 Punkte größer aus bei einer hohen Anzahl an Werks-Zitationen im Vergleich zu deren Nichtanzeige.

Auffallend ist hier der höchste aller Differenzwerte von −30,07 bei einem Vergleich der Kategorien *gering* und *hoch* bei UV 2 – Zitationszahl Werk auf der Stufe *gering* von UV 1 – Anzahl Downloads und analog zu den paarweisen Vergleichen von UV 3 * UV 2 * UV 1 in Abschnitt 4.4.3.1 ebenfalls auf der Stufe *keine Angabe* von UV 3 – Zitationszahl Autor. Auch dieser Vergleichswert stützt die Erwartung, dass eine hohe Anzahl Werks-Zitationen zu einer höheren Punktzahl bei der Bewertung führt als eine geringe Werks-Zitationen, ebenso in der Kombination *gering* und *hoch* bei UV 2 auf den Stufen von UV 1 – gering und UV 3 – gering, in der die Differenz −13,51 beträgt. Das Gegenteil ist der Fall bei zwei Kombinationen von UV 2, bei denen ein höherer Wert für die Stufe gering im Vergleich mit der Stufe hoch vorliegt. So weist der Vergleich von *gering* und *hoch* bei UV 1 – hoch und UV 3 – keine Angabe den Wert 21,21 auf, bei UV 1 – hoch und UV 3 – gering den Wert 12,85. Des Weiteren trifft erneut zu, dass UV 3 – keine Angabe die häufigsten Differenzwerte > 20 sowohl bei UV 2 – gering als auch bei UV 2 – hoch erzielt; UV 3 – hoch bewirkt keine, UV 3 – gering lediglich eine Differenz > 20.

In Abbildung 4.21 sind die Liniendiagramme über die Mittelwerte der Bewertungen von UV 2 – Zitationszahl Werk jeweils auf den Stufen von UV 1 – Anzahl Downloads und UV 3 – Zitationszahl Autor dargestellt. Im ersten Diagramm der oberen Reihe verlaufen die Linien der Kategorien von UV 2 bei einer geringen Anzahl Autoren-Zitationen auf den Stufen 1 – gering und 2 – hoch von UV 1 gegensätzlich: Die Linie *keine Angabe* hat einen leicht positiven Anstieg; die Linie *gering* zeigt kaum erkennbar einen ebenfalls leicht positiven Anstieg, wodurch der Eindruck eines annähernd parallelen Verlaufs entsteht. Die Linie der Kategorie *hoch* zeigt entgegen der Erwartung einen stark negativen Verlauf. Rechts von der Trennlinie weisen die Linien *gering* und *hoch* auf den Stufen 2 – hoch und 3 – keine Angabe von UV 1 einen stark positiven Anstieg auf, während die Linie *keine Angabe* negativ verläuft. Vergleicht man die Kategorien von UV 2 – Zitationszahl Werk auf den Stufen 1 – gering und 3 – keine Angabe von UV 1 – Anzahl Downloads, ist für die Kategorie *hoch* eine leicht positive Richtung,

für die Kategorie *keine Angabe* eine leicht negative Richtung erkennbar; für die Kategorie *gering* lässt sich hingegen ein positiver Anstieg feststellen.

Im ersten Diagramm der unteren Reihe verlaufen die Linien der Kategorien von UV 2 – Zitationszahl Werk bei einer geringen Anzahl Downloads (UV 1) auf den Stufen 1 – gering und 2 – hoch von UV 3 – Zitationszahl Autor ebenfalls ungleich: Die Linien *hoch* und *keine Angabe* verlaufen leicht negativ, wobei die Linie *hoch* einen stärken Abstieg aufweist als die Linie *keine Angabe*; die Linie *gering* weist einen positiven Anstieg auf. Auf den Stufen 2 – hoch und 3 – keine Angabe von UV 3 hingegen verlaufen die Linien *hoch* und *keine Angabe* positiv, die Linie *gering* stark negativ. Im Vergleich der Kategorien von UV 2 – Zitationszahl Werk auf den Stufen 1 – gering und 3 – keine Angabe von UV 1 – Anzahl Downloads lässt sich für *hoch* und *keine Angabe* ein leicht positiver Unterschied und für die Kategorie *gering* ein leicht negativer Unterschied feststellen.

Im zweiten Diagramm der oberen Reihe verlaufen die Linien der Kategorien *hoch* und *keine Angabe* von UV 2 – Zitationszahl Werk bei einer hohen Anzahl von Autoren-Zitationen (UV 3) auf den Stufen 1 – gering und 2 – hoch von UV 1 positiv, die Linie *gering* hingegen leicht negativ. Auf den Stufen 2 – hoch und 3 – keine Angabe (rechts von der Trennlinie) verläuft die Linie *hoch* ebenfalls positiv, die Linie *gering* zeigt einen leicht positiven Anstieg und die Linie *keine Angabe* verläuft stark negativ. Bei dem Vergleich der Kategorien von UV 2 auf den Stufen 1 – gering und 3 – keine Angabe von UV 1 – Anzahl Downloads wird deutlich, dass zwischen *gering* und *keine Angabe* ein kaum erkennbarer Unterschied vorliegt. Für die Kategorie *hoch* ist hingegen ein deutlich positiver Unterschied erkennbar, der nicht der Erwartung entspricht.

Ein ähnlich paralleler Verlauf der Linien *gering* und *keine* Angabe von UV 2 – Zitationszahl Werk auf den Stufen 1 – gering und 2 – hoch von UV 3 – Zitationszahl Autor ist bei einer hohen Anzahl Downloads (UV 1) (im zweiten Diagramm der unteren Reihe) erkennbar; die Linie *hoch* weist erwartungsgemäß einen deutlichen Anstieg auf. Auf den Stufen 2 – hoch und 3 – keine Angabe verlaufen die Linien *hoch* und *keine Angabe* nahezu parallel und erwartungsgemäß leicht negativ; die Linie *gering* verläuft positiv. Vergleicht man die Unterschiede der Kategorien von UV 2 auf den Stufen 1 – gering und 3 – keine Angabe von UV 3 – Zitationszahl Autor, lässt sich für *keine Angabe* nur ein vernachlässigbar geringer negativer Unterschied erkennen. Für die Kategorien *gering* und *hoch* zeigt sich ein deutlich positiver Anstieg, der den Erwartungen erneut widerspricht.

Auch im dritten Diagramm der oberen Reihe, also bei Nichtanzeige der Autoren-Zitationen, verlaufen die Linien der Kategorien von UV 2 – Zitationszahl Werk auf den Stufen 1 – gering und 2 – hoch von UV 1 – Anzahl Downloads ungleich: Die Linie *gering* zeigt einen starken positiven Anstieg, die Linie *hoch*

Tabelle 4.21 Mittelwerte für UV 3 * UV 1 * UV 2 – Paarweise Vergleiche

Stufen der Zitationszahl Autor (UV 3)	Stufen der Anzahl Downloads (UV 1)	Stufen der Zitationszahl Werk (UV 2) (I)	(J)	Differenz der Mittelwerte (I-J)	Std.-Fehler	p-Wert
gering	gering	gering	hoch	−13,51	1,50	< 0,001
			keine Angabe	−12,65	1,55	< 0,001
		hoch	keine Angabe	0,86	1,55	0,924
	hoch	gering	hoch	12,85	1,50	< 0,001
			keine Angabe	−15,00	1,49	< 0,001
		hoch	keine Angabe	−27,85	1,47	< 0,001
	keine Angabe	gering	hoch	2,28	1,54	0,361
			keine Angabe	11,76	1,48	< 0,001
		hoch	keine Angabe	9,48	1,60	< 0,001
hoch	gering	gering	hoch	9,76	1,57	< 0,001
			keine Angabe	1,61	1,52	0,643
		hoch	keine Angabe	−8,15	1,51	< 0,001
	hoch	gering	hoch	−0,09	1,54	1,000
			keine Angabe	−17,03	1,48	< 0,001
		hoch	keine Angabe	−16,95	1,48	< 0,001
	keine Angabe	gering	hoch	−8,99	1,45	< 0,001
			keine Angabe	4,78	1,47	0,004
		hoch	keine Angabe	13,77	1,45	< 0,001
keine Angabe	gering	gering	hoch	−30,07	1,42	< 0,001
			keine Angabe	−24,89	1,37	< 0,001
		hoch	keine Angabe	5,18	1,45	0,001
	hoch	gering	hoch	21,21	1,50	< 0,001
			keine Angabe	6,28	1,52	< 0,001
		hoch	keine Angabe	−14,93	1,52	< 0,001
	keine Angabe	gering	hoch	−4,53	1,51	0,008
			keine Angabe	5,57	1,50	0,001
		hoch	keine Angabe	10,10	1,54	< 0,001

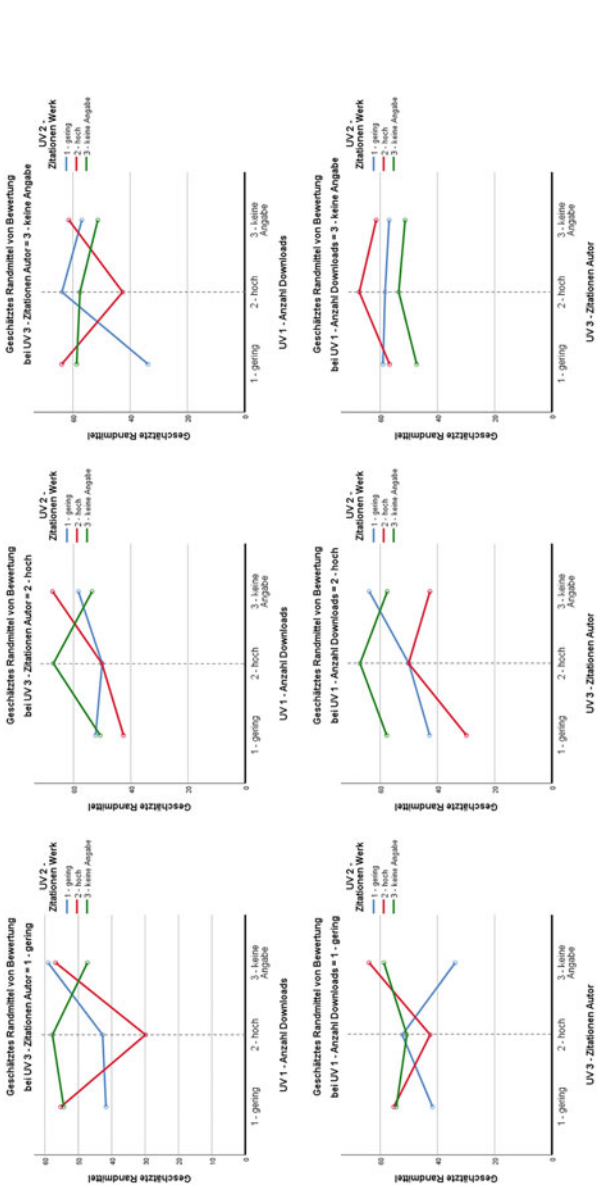

Abbildung 4.21 Mittelwerte der Bewertungen von UV 2 auf den Stufen von UV 1 bei UV 3 (obere Reihe) und auf den Stufen von UV 3 bei UV 1 (untere Reihe)

verläuft stark negativ und die Linie *keine Angabe* zeigt nur einen leicht negativen Verlauf. Ebenso gibt es einen parallelen Verlauf der Linien *gering* und *keine Angabe* auf den Stufen 2 – hoch und 3 – keine Angabe von UV 1 (rechts der Trennlinie), während die Linie *hoch* stark ansteigt. Die Kategorien *hoch* und *keine Angabe* lassen bei dem Vergleich der Stufen 1 – gering und 3 – keine Angabe von UV 1 einen leicht negativen Unterschied bei annähernder Parallelität erkennen, die Kategorie *gering* hingegen zeigt einen stark positiven Unterschied.

Im dritten Diagramm der unteren Reihe, also bei Nichtanzeige der Anzahl an Downloads, zeigen die Linien *hoch* und *keine Angabe* von UV 2 – Zitationszahl Werk auf den Stufen 1 – gering und 2 – hoch von UV 3 ebenfalls einen nur leichten Anstieg bei annähernder Parallelität; die Linie *gering* zeigt einen vernachlässigbar geringen negativen Verlauf. Rechts von der Trennlinie, also auf den Stufen 2 – hoch und 3 – keine Angabe von UV 3 – Zitationszahl Autor, verlaufen alle drei Linien annähernd gleich in negative Richtung, wobei die Linie *hoch* etwas stärker abfällt, während die Linien *gering* und *keine Angabe* nahezu parallel verlaufen. Deutlich parallel verhalten sich die Unterschiede zwischen den Kategorien hoch und keine Angabe auf den Stufen 1 – gering und 3 – keine Angabe von UV 3, für die Kategorie gering ist der Unterschied nur minimal in negativer Richtung.

4.4.3.3 Die Wirkung von UV 3 – Zitationszahl Autor auf den Stufen von UV 1 – Anzahl Downloads und UV 2 – Zitationszahl Werk

Tabelle 4.22 zeigt, dass lediglich fünf Differenzwerte nicht statistisch signifikant sind, während von den anderen Werten 18 statistisch signifikant sind bei $p < 0{,}001$ und vier Werte bei $p < 0{,}05$. Lediglich 6 Differenzwerte sind im Durchschnitt > 10, bei 3 Werten sind die Differenzen > 20. Diese 9 inhaltlich bedeutsamen Werte verteilen sich auf die Stufen von UV 2 und UV 1 folgendermaßen:

- Bei einer *geringen Anzahl* Werks-Zitationen (UV 2 – Stufe 1) liegt der größte Differenzenwert bei Stufe 2 – hoch von UV 1, also einer hohen Anzahl an Downloads, und gibt an, dass im Durchschnitt die Bewertung um 21,09 Punkte größer ausfällt, wenn die Anzahl der Autoren-Zitationen nicht angezeigt wird im Vergleich zu einer geringen Anzahl an Zitationen des Autors. Ebenfalls bei einer hohen Anzahl an Downloads ist der Differenzwert von 13,82 Punkten zu finden, der aussagt, dass die Bewertung um 13,82 Punkte größer ist bei Nichtanzeige der Autoren-Zitationen im Vergleich zu einer hohen Anzahl. Bei einer geringen Anzahl an Downloads (UV 1 – Stufe 1) fällt die Relevanzbewertung durchschnittlich um 18,50 Punkte größer aus, wenn die Anzahl der

Autoren-Zitationen hoch ist im Vergleich zu deren Nichtanzeige; dabei ist sie um 10,58 Punkte höher, wenn die Anzahl der Autoren-Zitationen hoch ist im Vergleich zu einer geringen Anzahl.

- Bei einer *hohen Anzahl* Werks-Zitationen (UV 2 – Stufe 2) ist die höchste Differenz von −21,33 Punkten bei der Relevanzbewertung auf Stufe 1 – gering von UV 1 – Anzahl Downloads zu finden. Dieser Wert gibt an, dass die Bewertung im Mittel um 21,33 Punkte höher ausfällt, wenn die Zitationszahl des Autors nicht angegeben ist im Vergleich mit einer hohen Anzahl Autoren-Zitationen. Ebenso ist bei einer geringen Anzahl Downloads (UV 1 – Stufe 1) die Punktzahl der Relevanzbewertung um 12,69 größer, wenn die Anzahl an Autoren-Zitationen gering ist im Vergleich zu einer hohen Anzahl Autoren-Zitationen. Bei einer hohen Anzahl Downloads (UV 1 – Stufe 2) liegt ein weiterer Differenzwert über der 20-Punkte-Schwelle. So ist die Bewertung um 20,20 Punkte größer, wenn die Autoren-Zitationen eine hohe Anzahl darstellen im Vergleich zu einer geringen Anzahl Autoren-Zitationen. Ebenfalls größer ist die Bewertung um 12,73 Punkte, wenn die Anzahl der Autoren-Zitationen nicht angegeben ist im Vergleich zu der Anzeige einer geringen Anzahl. Auf Stufe 3 – keine Angabe von UV 1 liegt lediglich ein Differenzwert über der 10-Punkte-Schwelle: Ist die Anzahl der Downloads nicht angegeben, fällt die Relevanzbewertung im Durchschnitt um 10,54 Punkte größer aus bei einer hohen Anzahl an Autoren-Zitationen im Vergleich zu einer geringen Anzahl Autoren-Zitationen.
- Bei *Nichtanzeige* der Werks-Zitationen (UV 2 – Stufe 3) gibt es weder einen Differenzwert > 20, noch einen Differenzwert > 10 bei den paarweisen Vergleichen von UV 3.

Im Vergleich mit den paarweisen Vergleichen in Tabelle 4.20 und Tabelle 4.21 sind hier die wenigsten Differenzwerte enthalten, die über der Schwelle zu den inhaltlich bedeutsamen Werten liegen. Zudem fällt auf, dass der höchste Differenzwert nur geringfügig größer ist als 20 (21,33). Allerdings sind auch hier Differenzwerte vorhanden, die bei einem Vergleich der Kategorien *gering* und *hoch* bei UV 3 – Zitationszahl Autor wie vermutet zeigen, dass eine hohe Anzahl Autoren-Zitationen zu einer höheren Punktzahl bei der Bewertung führt als eine geringe Autoren-Zitationen: Jeweils auf der Stufe *gering* von UV 1 und UV 2 (−10,58), bei der Stufe *hoch* von UV 1 und der Stufe *hoch* von UV 2 (−20,20) sowie bei der Stufe *keine Angabe* von UV 1 und UV 2 – hoch (−10,54) sind die Differenzwerte negativ. Positiv und somit entgegen der Erwartung ist der Differenzwert in der Kombination *gering* und *hoch* von UV 3 bei UV 1 – gering und UV 2 – hoch mit 12,69. Ferner sticht die Stufe *keine Angabe* von UV 2

Tabelle 4.22 Mittelwerte für UV 2 * UV 1 * UV 3 – Paarweise Vergleiche

Stufen der Zitationszahl Werk (UV 2)	Stufen der Anzahl Downloads (UV 1)	Stufen der Zitationszahl Autor (UV 3) (I)	(J)	Differenz der Mittelwerte (I-J)	Std.-Fehler	p-Wert
gering	gering	gering	hoch	−10,58	1,54	< 0,001
			keine Angabe	7,92	1,42	< 0,001
		hoch	keine Angabe	18,50	1,46	< 0,001
	hoch	gering	hoch	−7,27	1,52	< 0,001
			keine Angabe	−21,09	1,51	< 0,001
		hoch	keine Angabe	−13,82	1,51	< 0,001
	keine Angabe	gering	hoch	0,73	1,44	0,942
			keine Angabe	2,16	1,44	0,352
		hoch	keine Angabe	1,43	1,47	0,701
hoch	gering	gering	hoch	12,69	1,54	< 0,001
			keine Angabe	−8,64	1,51	< 0,001
		hoch	keine Angabe	−21,33	1,54	< 0,001
	hoch	gering	hoch	−20,20	1,52	< 0,001
			keine Angabe	−12,73	1,49	< 0,001
		hoch	keine Angabe	7,48	1,52	< 0,001
	keine Angabe	gering	hoch	−10,54	1,55	< 0,001
			keine Angabe	−4,65	1,61	0,012
		hoch	keine Angabe	5,90	1,49	< 0,001
keine Angabe	gering	gering	hoch	3,67	1,53	0,048
			keine Angabe	−4,33	1,50	0,012
		hoch	keine Angabe	−8,00	1,43	< 0,001
	hoch	gering	hoch	−9,30	1,44	< 0,001
			keine Angabe	0,20	1,50	0,999
		hoch	keine Angabe	9,50	1,48	< 0,001
	keine Angabe	gering	hoch	−6,25	1,50	< 0,001
			keine Angabe	−4,03	1,54	0,026
		hoch	keine Angabe	2,22	1,50	0,361

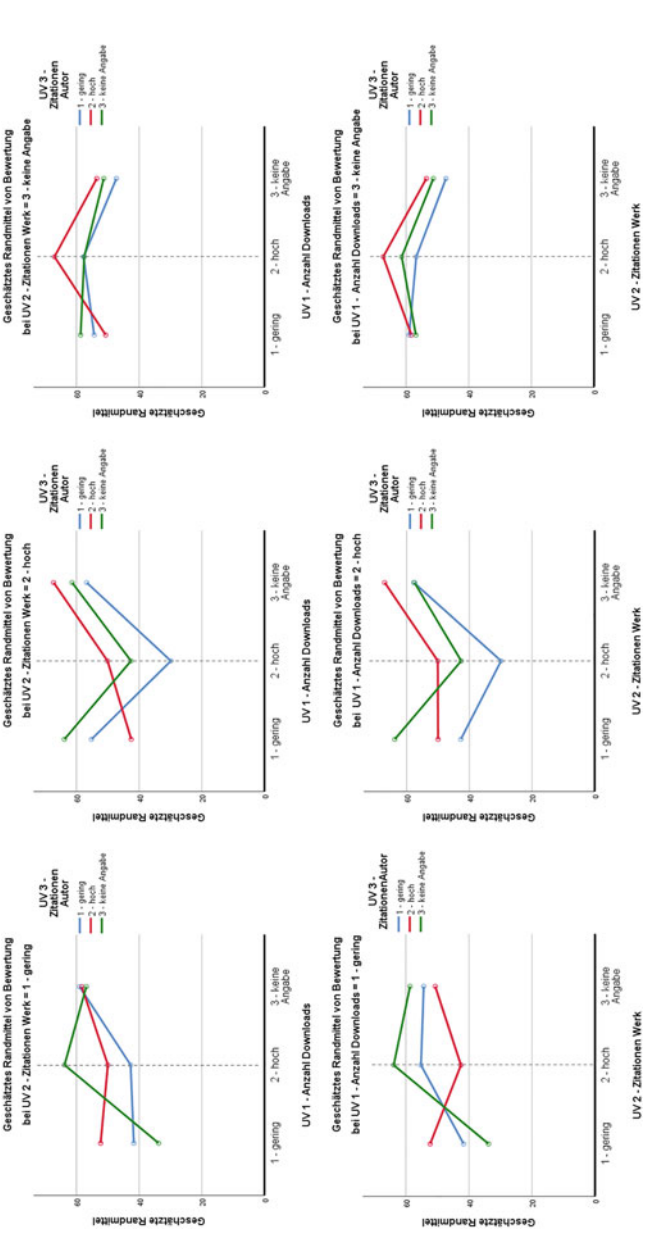

Abbildung 4.22 Mittelwerte der Bewertungen von UV 3 auf den Stufen von UV 1 bei UV 2 (obere Reihe) und auf den Stufen von UV 2 bei UV 1 (untere Reihe)

hervor, da hier bei dem Vergleich der Werte auf den Stufen von UV 3 keinerlei Differenzwerte > 10 vorhanden sind.

Abbildung 4.22 zeigt die Liniendiagramme der durchschnittlichen Bewertungen von UV 3 – Zitationszahl Autor auf den jeweiligen Stufen von UV 1 – Anzahl Downloads und UV 2 – Zitationszahl Werk. Im ersten Diagramm der oberen Reihe verlaufen die Linien bei einer geringen Anzahl von Werkszitationen auf den Stufen 1 – gering und 2 – hoch von UV 1 erneut gegensätzlich: Die Linie *gering* zeigt einen leicht positiven Anstieg, die Linie *hoch* einen leicht negativen Abstieg und die Linie *keine Angabe* einen stark positiven Anstieg. Auf den Stufen 2 – hoch und 3 – keine Angabe von UV 1 (rechts von der Trennlinie) hingegen verläuft die Linie *keine Angabe* leicht negativ, die Linie *hoch* leicht positiv und die Linie *gering* stark positiv. Bei einem Vergleich der Kategorien von UV 3 auf den Stufen 1 – gering und 3 – keine Angabe von UV 1 – Anzahl Downloads ist für alle ein positiver Unterschied erkennbar, wobei alle relativ nah beieinander liegen und der größte Unterschied für die Kategorie *keine Angabe* besteht. Das heißt, dass bei Nichtanzeige der Anzahl von Downloads (UV 1) und einer geringen Anzeige von Werks-Zitationen (UV 2) die durchschnittlichen Relevanzbewertungen von UV 3 – Zitationszahl Autor zum einen nur einen sehr geringen Abstand zueinander aufweisen und zum anderen alle entgegen der Erwartung höher ausfallen als bei einer geringen Anzahl Downloads.

Im ersten Diagramm der unteren Reihe ähnelt der Verlauf der Linien der Kategorien von UV 3 – Zitationszahl Autor sehr dem Linienverlauf im ersten Diagramm der oberen Reihe links von der Trennlinie: Die Linie *keine Angabe* zeigt auf den Stufen 1 – gering und 2 – hoch von UV 2 – Zitationszahl Werk bei einer geringen Anzahl Downloads (UV 1) einen stark positiven Anstieg, die Linie *gering* einen positiven Anstieg und die Linie *hoch* verläuft negativ. Rechts von der Trennlinie, also auf den Stufen 2 – hoch und 3 – keine Angabe von UV 2 – Zitationszahl Werk, verlaufen die Linien *gering* und *keine Angabe* nahezu gleich leicht negativ, wobei der Unterschied in der Kategorie *gering* von UV 3 – Zitationszahl Autor kaum zu erkennen ist; die Linie *hoch* zeigt hier erneut einen leicht positiven Anstieg. Bei dem Vergleich der Kategorien von UV 3 – Zitationszahl Autor auf den Stufen 1 – gering und 3 – keine Angabe von UV 2 – Zitationszahl Werk ist für die Kategorie hoch ein leicht negativer Abstieg, für die Kategorie *gering* ein leicht positiver und für die Kategorie *keine Angabe* ein etwas stärkerer positiver Unterschied erkennbar.

Im zweiten Diagramm der oberen Reihe verlaufen bei einer hohen Anzahl Werkszitationen (UV 2) auf den Stufen 1 – gering und 2 – hoch von UV 1 – Anzahl Downloads erneut zwei Linien der Kategorien von UV 3 – Zitationszahl Autor ähnlich: Die Linien *gering* und *keine Angabe* verlaufen stark negativ bei

annähernder Parallelität, während die Linie *hoch* einen leichten positiven Anstieg zeigt. Dieser nahezu parallele Verlauf setzt sich auf den Stufen 2 – hoch und 3 – keine Angabe von UV 1 – Anzahl Downloads bei einem stark positiven Anstieg fort, wobei hier auch die Linie *hoch* parallel zu den anderen verläuft. Auch bei dem Vergleich der Kategorien *gering* und *keine Angabe* auf den Stufen 1 – gering und 3 – keine Angabe von UV 1 wird ein paralleler Verlauf ersichtlich, wobei jedoch die Kategorie *gering* einen kaum wahrnehmbaren Unterschied aufweist, während für die Kategorie *keine Angabe* der Unterschied minimal gering ausfällt; die Kategorie hoch hingegen zeigt einen großen Unterschied in positive Richtung, was nicht der Erwartung entspricht.

Im zweiten Diagramm der unteren Reihe verlaufen die Linien der Kategorien *gering* und *keine Angabe* von UV 3 – Zitationszahl Autor auf den Stufen 1 – gering und 2 – hoch von UV 2 – Zitationszahl Werk ähnlich negativ wie im zweiten Diagramm der oberen Reihe auf den Stufen 1 – gering und 2 – hoch von UV 1 – Anzahl Downloads; die Linie hoch verläuft dagegen weder positiv noch negativ. Auf den Stufen 2 – hoch und 3 – keine Angabe von UV 2 (rechts von der Trennlinie) verlaufen erneut die Linien *hoch* und *keine Angabe* parallel bei positivem Anstieg, die Linie *gering* verläuft ebenfalls positiv bei einem stärkeren Anstieg. Bei dem Vergleich der Kategorien von UV 3 – Zitationszahl Autor auf den Stufen 1 – gering und 3 – keine Angabe von UV 2 – Zitationszahl Werk kann ein nahezu paralleler Verlauf für *gering* und *hoch* festgestellt werden, wobei auch hier die Werte auf der Stufe 3 – keine Angabe von UV 2 höher liegen als auf Stufe 1 – gering. Für die Kategorie *keine Angabe* ist ein geringer Unterschied in negativer Richtung erkennbar.

Im dritten Diagramm der oberen Reihe verlaufen die Linien der Kategorien *gering* und *keine Angabe* von UV 3 – Zitationszahl Autor bei Nichtanzeige der Werks-Zitationen (UV 2) auf den Stufen 1 – gering und 2 – hoch von UV 1 – Anzahl Downloads zwar sehr nah beieinander, allerdings gegensätzlich: Die Linie *gering* verläuft leicht positiv, die Linie *keine Angabe* leicht negativ. Die Linie *hoch* hat dagegen einen deutlich positiven Anstieg. Rechts von der Trennlinie, also auf den Stufen 2 – hoch und 3 – keine Angabe, verlaufen alle drei Linien gleich negativ, wobei die Linie *keine Angabe* den geringsten Abstieg zeigt und die Linien *gering* und *hoch* nahe parallel verlaufen. Ebenfalls gleich und parallel zeigt sich der Unterschied in den Kategorien *gering* und *keine Angabe* bei dem Vergleich der Stufen 1 – gering und 3 – keine Angabe von UV 1 – Anzahl Downloads, während der Unterschied für die Kategorie *hoch* lediglich minimal positiv ausfällt.

Im dritten Diagramm der unteren Reihe zeigt sich ein ähnliches Ergebnismuster wie in dem dritten Diagramm der oberen Reihe: Bei Nichtanzeige der

Downloads (UV 1) verlaufen die Linien der Kategorien hoch und keine Angabe von UV 3 – Zitationszahl Autor auf den Stufen 1 – gering und 2 – hoch von UV 2 – Zitationszahl Werk leicht positiv, die Linie gering leicht negativ. Auf den Stufen 2 – hoch und 3 – keine Angabe (rechts von der Trennlinie) zeigt sich auch hier ein annähernd paralleler Verlauf aller drei Linien in erwartungskonform negativer Richtung. Die Unterschiede der Kategorien von UV 3 – Zitationszahl Autor auf der Stufe 1 – gering von UV 2 – hoch im Vergleich zur Stufe 2 – hoch von UV 2 – Zitationszahl Werk sind erneut nur minimal in negativer Richtung; zwischen den Kategorien *gering* und *hoch* kann bei großzügiger Betrachtung von einem parallelen Verlauf ausgegangen werden.

Sowohl bei den Haupteffekten als auch bei den Interaktionseffekten der 1. Ordnung konnte ein komplexes Ergebnismuster festgestellt werden, das den Erwartungen nicht entspricht. Die Ergebnisse der 2-fach-Interaktionen zeigen, dass sich dieses komplexe Muster bei dem Vergleich aller unabhängigen Variablen untereinander auf den drei Stufen *gering – hoch – keine Angabe* fortsetzt.

Betrachtet man ausschließlich die als inhaltlich bedeutsam zu beurteilenden Differenzwerte der paarweisen Vergleiche, die einen Unterschied von mindestens 10 Punkten und 20 Punkten aufweisen, fällt auf, dass die beiden höchsten Differenzwerte > 30 beide auf der Stufe 3 – keine Angabe von UV 3 – Zitationszahl Autor liegen: Ist zusätzlich die Zitationszahl Werk gering (UV 2 – Stufe 1), ist die Bewertung durchschnittlich 30,02 Punkte höher, wenn die Anzahl der Downloads hoch (UV 1 – Stufe 2) ist im Gegensatz zu einer geringen Anzahl (UV 1 – Stufe 1) (vgl. Tabelle 4.20); ist dagegen zusätzlich die Anzahl der Downloads gering (UV 1 – Stufe 1), fällt die Bewertung im Durchschnitt um 30,07 Punkte höher aus, wenn die Zitationszahl Werk hoch (UV 2 – Stufe 2) ist im Vergleich zu einer geringen Anzahl (UV 2 – Stufe 1) (vgl. Tabelle 4.21).

Sehr ähnlich sind zudem die Differenzwerte im Vergleich der Stufen 1 – gering mit 2 – hoch jeweils für UV 1 und UV 2: Die Durchschnittsbewertung ist um 23,07 Punkte höher bei Nichtanzeige der Anzahl Downloads (UV 1 – Stufe 3) im Vergleich mit einer geringen Anzahl Downloads (UV 1 – Stufe 1) (vgl. Tabelle 4.20), um 24,89 Punkte ist sie höher bei Nichtanzeige der Zitationszahl Werk (UV 2 – Stufe 3) im Vergleich mit einer geringen Anzahl Werkszitationen (UV 2 – Stufe 1) (vgl. Tabelle 4.21). Ein ebenso ähnliches und den Erwartungen nicht entsprechendes Bild zeigt sich beispielsweise auf der Stufe 2 – hoch jeweils für UV 1 und UV 2 bei Nichtanzeige der Autorenzitationen (UV 3 – Stufe 3): Die Bewertung ist um 21,26 Punkte im Durchschnitt kleiner bei einer hohen Anzahl Downloads im Vergleich zu einer geringen Anzahl Downloads, wenn die Anzahl der Werkszitationen hoch ist (vgl. Tabelle 4.20); um 21,21 Punkte ist sie kleiner bei einer hohen Anzahl Werkszitationen im Vergleich mit einer geringen Anzahl

Werkszitationen, wenn die Anzahl der Downloads hoch ist (vgl. Tabelle 4.21). Die Besonderheit von Kategorie 3 – keine Angabe wird hier erneut deutlich.

4.5 Diskussion der Ergebnisse im Kontext der Studienmethodik

Dieser Abschnitt beginnt mit einer Zusammenfassung des Vorgehens bei der Entwicklung des experimentellen Untersuchungsdesigns (vgl. Abschnitt 4.1), der Datenerhebung (vgl. Abschnitt 4.2) und der Datenauswertung (vgl. Abschnitt 4.3). Anschließend werden die in Abschnitt 4.4 berichteten Ergebnisse diskutiert. Dabei werden mögliche Gründe für die Uneindeutigkeit der statistischen Ergebnisse erörtert. Zu beachten ist hierbei, dass aufgrund fehlender vergleichbarer Studien zur Erforschung von Relevanzkriterien anhand eines experimentellen Designs (vgl. Abschnitt 2.2.4) keine Erkenntnisse anderer Untersuchungen in die hier vorgenommene Ergebnisdiskussion einfließen können.

Die Begründung für die Entwicklung und Durchführung eines experimentellen Designs wurde bereits im Zusammenhang mit der Entwicklung der Forschungsfrage F1 in Abschnitt 2.3 diskutiert. Das Hauptargument liegt hierbei darin, unter Berücksichtigung von Manipulation und Kontrolle kausale Schlussfolgerungen über den Zusammenhang zwischen Ursache und Wirkung zweier Variablen ableiten zu können. Der Vorteil eines mehrfaktoriellen Designs gegenüber mehreren einfaktoriellen Untersuchungen besteht in der Feststellung von Interaktionen, die Abhängigkeiten von den verschiedenen Ausprägungen der einzelnen unabhängigen Variablen sichtbar machen. Bedenkt man die Komplexität von Relevanz und die diversen Elemente eines Suchergebnisses, die als Relevanzmerkmale dienen können und durch ihr Zusammenwirken die Relevanzbewertung beeinflussen, wäre ein einfaktorielles Design wenig zielführend.

Die in diesem Experiment untersuchten unabhängigen Variablen UV 1 – *Anzahl Downloads*, UV 2 – *Zitationszahl Werk* und UV 3 – *Zitationszahl Autor* mit jeweils drei Ausprägungen (gering – hoch – keine Angabe) stellen die operationalisierten Relevanzkriterien für Popularität bzw. im Fall der *Zitationszahl Autor* für Autorität dar. Ihr Einfluss auf die Relevanzbewertung von Suchergebnissen (Surrogaten) in akademischen Suchsystemen als zu messende abhängige Variable wird als positiv angenommen. Die abhängige Variable *Bewertung* wurde operationalisiert als Punktzahl der Bewertung über die Nützlichkeit des Surrogates hinsichtlich eines Informationsbedürfnisses, die mithilfe einer Schieberegler-Skala mit 101 Abstufungen durch die Versuchspersonen explizit angezeigt wurde. Konkret bestand die Annahme, dass eine höhere Anzahl an

Downloads oder Zitationen eines Werks bzw. eines Autors mit den Bewertungen der Suchergebnisse positiv korreliert.

Die Entscheidung für ein Within-Subjects-Designs und gegen ein Between-Subjects-Design beruhte zum einen auf inhaltlichen Gründen hinsichtlich des Zusammenhangs der Variablen und zum anderen auf dem Argument, einen höheren Stichprobenumfang erreichen zu können. Zudem stellten personengebundene Störvariablen kein Problem dar, weil diese in einem Within-Subjects-Design vollständig parallelisiert und mögliche unerwünschte Effekte über alle Bedingungen ausgeglichen sind.

Das mehrfaktorielle Within-Subjects-Design wurde vollständig als Online-Fragebogen umgesetzt, d. h. die Versuchspersonen wurden allen 27 möglichen Bedingungskombinationen der drei UVn auf jeweils drei Stufen gleichermaßen ausgesetzt. Die Reihenfolge der angezeigten Bedingungen wurde randomisiert, um möglichen Reihenfolge- und Positionseffekten vorzubeugen. Eine Bedingung bestand in der Präsentation eines Surrogates, das die manipulierten Popularitätsdaten enthält. Die insgesamt 27 Surrogate wurden auf 3 nacheinander zu bearbeitenden Aufgaben zu den Themen *Altmetrics*, *Peer Review* und *Wikipedia* verteilt, wobei nicht nur die Reihenfolge der Surrogate innerhalb einer Aufgabe, sondern auch die Reihenfolge der drei Aufgaben im Online-Fragebogen randomisiert wurde. Eine Aufgabe beinhaltete eine kurze Situationsbeschreibung gefolgt von der Beschreibung eines Informationsbedürfnisses, zu dem die gelisteten Suchergebnisse in Hinblick auf deren Nützlichkeit zur Befriedigung des beschriebenen Bedürfnisses bewertet werden sollten. Die Entwicklung dieser Beschreibungstexte, die auch als Vignetten bezeichnet werden können, orientierte sich an dem im Interactive Information Retrieval häufig verwendeten Konzept der *Simulated Work Task Situation.* Für die Erstellung der Surrogate wurde anhand zuvor entworfener Auswahlkriterien auf real existierende Dokumentsurrogate zurückgegriffen. Das Ergebnis der Fragebogenkonstruktion stellt ein multifaktorielles Online-Survey (auch Online-Vignettenanalyse) dar, das im Gegensatz zur oft als zu künstlich empfundenen Laborsituation in einem realen Umfeld der Versuchspersonen durchgeführt wird. Da hier dennoch „die experimentelle Situation in hohem Ausmaß kontrolliert werden kann“ (Berger & Wolbring, 2015, S. 46), darf von einer höheren externen Validität ausgegangen werden.

Um die mithilfe des Statistik-Tools *G*Power* a-priori berechnete optimale Stichprobengröße von $n = 577$ zu erreichen, wurden mehr als 15.000 Wissenschaftliche Mitarbeiterinnen und Mitarbeiter sowie (Post-)Doktorandinnen und (Post-)Doktoranden unterschiedlicher Fachrichtungen an verschiedenen Universitäten Deutschlands zur Teilnahme an der Studie per E-Mail eingeladen. Über den wahren Zweck des Experiments wurden die Teilnehmenden nicht vor

Beginn, sondern am Ende der Befragung aufgeklärt, um eine mögliche unerwünschte Beeinflussung auf die erhobenen Daten zu verhindern. Stattdessen wurde die Umfrage unter dem Titel „Teilnehmende für Online-Umfrage zur Nutzung wissenschaftlicher Suchsysteme“ beworben.

Die in einem Zeitraum von 36 Tagen im Sommer 2019 erhobenen Daten wurden aufbereitet, bereinigt und mit SPSS 25 einer Mehrebenenanalyse unterzogen. Mithilfe der Mehrebenenanalyse können die Wirkungen der einzelnen unabhängigen Variablen (Haupteffekte) analysiert sowie die Abhängigkeiten dieser einzelnen Effekte von den Stufen der jeweils anderen unabhängigen Variablen (Interaktionseffekte) geprüft werden.

Das Experiment zur Untersuchung des Einflusses von Popularitätsdaten auf die Relevanzbewertung von Suchergebnissen in akademischen Suchsystemen führte nicht zu den – wie angenommen – eindeutigen Ergebnissen. Dass die *inhaltlichen* Hypothesen nicht bestätigt werden können, wurde bereits an mehreren Stellen erwähnt. Eine *statistische* Hypothesenprüfung entfällt aufgrund der fehlenden statistischen Hypothesen über konkrete Erwartungen zur Effektgröße, denn für deren Aufstellung hätten statistische Ergebnisse bzw. statistische Parameter aus vergleichbaren Studien herangezogen werden müssen. Da diese jedoch weder für die Haupteffekte noch für die Interaktionseffekte vorlagen, wurden lediglich inhaltliche bzw. empirische Hypothesen über die Haupteffekte formuliert. Diese beschreiben einen positiven Einfluss der UV 1, UV 2 und UV 3 auf die AV (vgl. Abschnitt 4.1.3):

- H1: Die Downloadhäufigkeit eines Werks hat einen positiven Einfluss auf die Relevanzbewertung.
- H2: Die Zitationshäufigkeit eines Werkes hat einen positiven Einfluss auf die Relevanzbewertung.
- H3: Die Zitationshäufigkeit des Autors hat einen positiven Einfluss auf die Relevanzbewertung.

Sowohl die Haupteffekte als auch die Interaktionseffekte 1. Ordnung (2-fach-Interaktionen) und 2. Ordnung (3-fach-Interaktionen) sind statistisch signifikant. Das bedeutet, dass der Nachweis des Effekts der jeweils einzelnen unabhängigen Variablen aufgrund der statistisch signifikanten Wechselwirkungen mit den anderen UVn nicht (mehr) standhalten kann. Dadurch sind die Haupteffekte – exklusiv betrachtet – wenig aussagekräftig. Stattdessen stellt die 3-fach-Interaktion das zentrale Ergebnis der gesamten Mehrebenenanalyse dar.

Die Schwierigkeit der Interpretation der Daten besteht unter anderem aufgrund der zahlreichen paarweisen Vergleiche, die für die 3-fach-Interaktion analysiert

wurden. Die hohe Anzahl an paarweisen Vergleichen ergibt sich aus der Anzahl der unabhängigen Variablen und der Anzahl der Stufen. Da in dem Experiment drei unabhängige Variablen mit jeweils drei Stufen untersucht wurden, liegen die Daten aus 27 verschiedenen Bedingungen vor, die für die 3-fach-Interaktion für jede UV auf allen Stufen der jeweils anderen beiden UVn auszuwerten sind.

Damit die statistischen Ergebnisse sinnvoll interpretiert werden können, wurde der Versuch unternommen, diese in Bezug auf ihre inhaltliche Relevanz einzugrenzen. So wurde der Fokus auf diejenigen Ergebnisse der paarweisen Vergleiche gelegt, die Unterschiede zwischen den Bewertungen von mindestens 10 bzw. 20 Punkten legten. Im Zusammenspiel der drei unabhängigen Variablen zeigte sich, dass die beiden werksbezogenen Variablen, d. h. die Anzahl der Downloads (UV 1) und die Anzahl der Zitationszahl eines Werks (UV 2), im Vergleich zur Anzahl der Zitationen eines Autors (UV 3), eine größere Anzahl dieser inhaltlich bedeutsamen Unterschiede aufweisen, von denen jeweils ein Wert sogar über 30 Punkte liegt (vgl. Tabelle 4.19). Die Zitationszahl Werk liegt hierbei mit insgesamt 14 inhaltlich bedeutsamer Differenzwerte vor der Anzahl der Downloads mit 13 Werten und vor der Zitationszahl des Autors mit 9 Werten. Dies mag zu der Schlussfolgerung verleiten, die Zitationszahl Werk (UV 2) von den drei unabhängigen Variablen als diejenige mit dem größten Effekt auf die Relevanzbewertung zu beurteilen; dies lässt sich anhand der statistischen Ergebnisse insbesondere bei der 3-fach-Interaktion jedoch nicht eindeutig nachweisen.

Insbesondere die Differenzwerte, die auf Stufe 3 – keine Angabe hervorgerufen wurden, weichen von den im Rahmen der Hypothesenformulierung erläuterten Annahmen deutlich ab. Vermutet wurde, dass (a) bei Stufe 2 – hoch die Punktzahl der Relevanzbewertung im Durchschnitt größer als bei Stufe 1 – gering oder Stufe 3 – keine Angabe wäre und (b) bei Stufe 3 – keine Angabe die Punktzahl der Relevanzbewertung im Durchschnitt kleiner als bei Stufe 1 – gering oder Stufe 2 – hoch wäre. Die Hypothesen lassen demzufolge eine Ordinalskala vermuten, obwohl es sich tatsächlich um eine Nominalskala handelt, da den drei Kategorien ihre Werte aus einem bestimmten Wertebereich zugewiesen wurden. Diese vermutete Rangfolge der Kategorien lässt sich anhand der statistischen Ergebnisse nicht bestätigen. Es ist nicht der Fall, dass die Differenzwerte bei den paarweisen Vergleichen auf Stufe 3 – keine Angabe immer niedriger sind als die Werte auf Stufe 1 – gering oder Stufe 2 – hoch. Es ist auch nicht der Fall, dass die Differenzwerte auf Stufe 2 – hoch immer höher sind als die Werte auf Stufe 1 – gering oder Stufe 3 – keine Angabe.

Die Eingrenzung der Differenzwerte auf inhaltlich bedeutsame Unterschiede von mindestens 10 Punkten hat diesbezüglich keine klareren Erkenntnisse bewirkt; dennoch kann sie als sinnvolle Maßnahme für die Interpretation der

Ergebnisse bewertet werden, weil diese Eingrenzung mit der Beschäftigung mit dem Unterschied zwischen Ergebnissen, die statistisch signifikant (statistisch bedeutsam) sind und Ergebnissen, die inhaltlich bedeutsam sind, einherging. In diesem Zusammenhang schreiben Döring & Bortz:

> Häufig wird ein statistisch signifikantes Ergebnis automatisch für ein bedeutsames Ergebnis gehalten, insbesondere wenn es auf einer großen Stichprobe basiert. Gerade bei großen Stichproben können jedoch auch sehr kleine, praktisch unbedeutende Effekte statistisch signifikant werden. (2016, S. 668)

Auch bei einer Aufteilung der Gesamtstichprobe in die sechs fachdisziplinspezifischen Gruppen (vgl. Abschnitt 4.4.1) und deren statistischer Mehrebenenanalyse zeigt sich kein klareres Bild. Hier liegen zwar ebenfalls statistisch signifikante Ergebnisse, aber auch nicht statistisch signifikante Ergebnisse vor, die über die Richtung der Effekte keine eindeutigen Aussagen zulassen. Die Haupteffekte sowie die Interaktionseffekte 2. Ordnung für jede der sechs Fachdisziplinen sind den Tabellen der Mittelwerte und paarweisen Vergleiche in Anhang 4.2 im elektronischen Zusatzmaterial zu entnehmen. Bei der Betrachtung dieser statistischen Ergebnisse ist zu beachten, dass die jeweilige Anzahl der analysierten Fälle und Versuchspersonen (zu einer Versuchsperson liegen jeweils 27 Fälle, d. h. bewertete Surrogate, vor) zwischen den Gruppen sehr stark variiert, sodass ein direkter Gruppenvergleich untereinander sowie ein Vergleich einer Gruppe aus einer verhältnismäßig geringen Anzahl an Versuchspersonen (z. B. Humanmedizin mit $n = 27$ bzw. 4,3 % der Gesamtstichprobe) mit den Ergebnissen der Gesamtstichprobe nicht sinnvoll ist. Neben den beiden größten Gruppen der Naturwissenschaften ($n = 264$ bzw. 42,1 % der Gesamtstichprobe) und Sozialwissenschaften ($n = 155$ bzw. 24,7 % der Gesamtstichprobe) zeigen die restlichen vier einen jeweils kleineren Stichprobenumfang mit $n < 100$. Auf die erneute Mehrebenenanalyse unter Hinzunahme der anderen erhobenen demografischen Variablen, wie Alter und Geschlecht, in das Mehrebenenmodell wurde verzichtet, um das ohnehin komplexe Ergebnismuster nicht weiter zu verkomplizieren.

Die Tatsache, dass die statistischen Ergebnisse kein eindeutiges Bild über den Einfluss der untersuchten Popularitätsdaten auf die Relevanzbewertung von Surrogaten im akademischen Kontext liefern, kann unter anderem mit der Variation der unabhängigen Variablen auf Stufe 3 als *keine Angabe* begründet werden.

Motiviert war die Hinzunahme der Ausprägung *keine Angabe* zu den drei unabhängigen Variablen zum einen damit, dass diese die Präsentation von Suchergebnissen in traditionellen akademischen Suchsystemen, d. h. ohne integrierte zusätzliche Daten wie Popularitätsdaten, wiedergeben; zum anderen wurden die

Bewertungen des Surrogats, das die Bedingung, in der alle UVn auf Stufe 3 – keine Angabe variiert sind (S333), repräsentiert, als „Baseline“ für die Bewertung der thematischen Relevanz erachtet (Abbildung 4.23). Zunächst hebt sich diese Kategorie von den anderen beiden Kategorien *gering* und *hoch* dahingehend ab, dass ihr kein quantitativer Wert zugewiesen wurde. Daher stellt sich die Frage, an welchen Elementen des Surrogats sich die Versuchspersonen orientierten, wenn für die UV kein Wert angegeben war, sondern das Kürzel n. a. (*not available*) angezeigt wurde. In der Erläuterung zu den Aufgaben im Fragebogen wurde ausdrücklich darauf hingewiesen, dass dieses Kürzel nicht gleichbedeutend ist mit Null, sondern das Fehlen eines Wertes kennzeichnet. Ob jede VPn den Erläuterungstext aufmerksam gelesen und diesen Hinweis tatsächlich berücksichtigt hat, ist ungewiss.

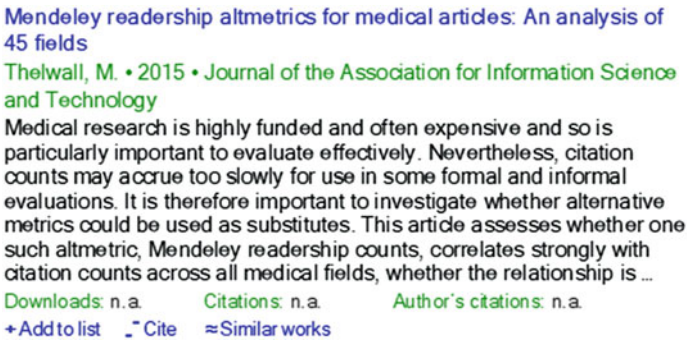

Abbildung 4.23 Surrogat mit der Bedingung S333 aus Aufgabe 1 – Altmetrics

Im Zusammenhang mit dem Aspekt der thematischen Relevanz ist die Auswahl der Surrogate in Hinblick auf die Beschreibung der Informationsbedürfnisse kritisch zu betrachten. Es stellte sich heraus, dass das Surrogat S221 im Durchschnitt die geringste Anzahl an Bewertungspunkten (29,87) erzielte (vgl. Tabelle 4.18 und Abbildung 4.19), obwohl die Kombination der Stufen *hoch* – *hoch* – *gering* einen höheren Wert vermuten lassen würde; der Grund ist vermutlich der, dass das Basiskriterium der thematischen Relevanz für die VPn ausschlaggebend war für die Bewertung, dieses jedoch als verhältnismäßig gering erachtet wurde. Dies ist bei einer eingehenden Prüfung des Surrogats (Abbildung 4.24) im Zusammenhang mit der Beschreibung des Informationsbedürfnisses durchaus nachvollziehbar; der Beschreibungstext bzw. die Aufgabe lauteten:

Viele Menschen nutzen die Online-Enzyklopädie Wikipedia – die deutschsprachige Webseite wird eigenen Angaben zufolge täglich Millionen Mal aufgerufen. Trotz ihrer Beliebtheit wird Wikipedia im Bildungskontext und im Hochschulbereich gemeinhin nicht als zitierfähige Informationsquelle erachtet, da Zweifel an der Güte bzw. Qualität von Wikipedia-Artikeln bestehen.
Ihr Informationsbedürfnis: Sie möchten herausfinden, ob diese Zweifel in Hinblick auf Wikipedia und Lehre berechtigt sind.

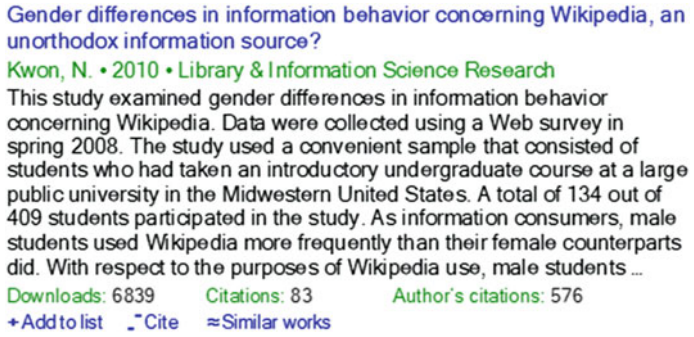
Gender differences in information behavior concerning Wikipedia, an unorthodox information source?
Kwon, N. • 2010 • Library & Information Science Research
This study examined gender differences in information behavior concerning Wikipedia. Data were collected using a Web survey in spring 2008. The study used a convenient sample that consisted of students who had taken an introductory undergraduate course at a large public university in the Midwestern United States. A total of 134 out of 409 students participated in the study. As information consumers, male students used Wikipedia more frequently than their female counterparts did. With respect to the purposes of Wikipedia use, male students …
Downloads: 6839 Citations: 83 Author's citations: 576
+Add to list Cite ≈Similar works

Abbildung 4.24 Surrogat mit der geringsten Durchschnittsbewertung aus Aufgabe 3 – Wikipedia

Die Aufnahme dieses Surrogats als eines der zu bewertenden Suchergebnisse war demzufolge ein Fehler, der vermutlich auf das Vorgehen bei der Auswahl der Surrogate als ersten Schritt und der darauffolgenden Entwicklung der Informationsbedürfnisse zurückzuführen ist: Die ausgewählten Surrogate behandeln zwar alle das Thema Wikipedia, aber die Beschreibung des Informationsbedürfnisses in Hinblick auf den Kontext des Hochschulbereichs bzw. der Lehre trifft auf das betreffende Surrogat (Titel und Abstract) weniger zu.

Um einer Diskrepanz bei der Auswahl der Surrogate und der Formulierung der Informationsbedürfnisbeschreibungen in Hinblick auf die thematische Übereinstimmung (*Aboutness*), anhand derer die thematische Relevanz abgeleitet werden kann, vorzubeugen, hätte die Übereinstimmung der *Aboutness* der Surrogate mit den Beschreibungen der Informationsbedürfnisse gesondert durch unabhängige Dritte, also Personen mit einem bibliotheks- oder informationswissenschaftlichen Hintergrund, überprüft werden müssen.

In diesem Zusammenhang liegt es nahe, nicht nur Unterschiede bei den Bewertungen einzelner Surrogate genauer zu betrachten, sondern auch zu prüfen, ob

Auffälligkeiten bei den Bewertungen zwischen den Aufgaben bestehen. Ein varianzanalytischer Vergleich der durchschnittlichen Bewertungen aller Surrogate innerhalb einer Aufgabe mit den durchschnittlichen Bewertungen der Surrogate der jeweils anderen Aufgaben zeigt, dass statistisch signifikante Unterschiede bei $p < 0{,}001$ zwischen den Bewertungen der drei Aufgaben existieren: So wurden die Surrogate von Aufgabe 1 von allen Versuchspersonen im Durchschnitt mit 57,3 Punkten bewertet, die Surrogate von Aufgabe 2 mit 53,6 Punkten und die Surrogate der Aufgabe 3 wurden mit durchschnittlich 47,5 Punkten am geringsten bewertet (Abbildung 4.25). Die Mittelwerte und paarweisen Vergleiche sowie die Tafel der Varianzanalyse sind enthalten in Anhang 4.3 im elektronischen Zusatzmaterial. Eine Aufnahme der Aufgabenzugehörigkeit als weiterer Faktor in das statistische Mehrebenenmodell wäre nur möglich, wenn die experimentellen Bedingungen in jeder Aufgabe dieselben wären. Dies ist hier jedoch nicht der Fall, denn die insgesamt 27 experimentellen Bedingungen wurden über die 3 Aufgaben zufällig verteilt, sodass in jeder Aufgabe jeweils 9 verschiedene Bedingungen vorlagen.

Die statistisch signifikanten Unterschiede der Bewertungen zwischen den Aufgaben deuten zum einen auf einen inhaltlichen Effekt der jeweiligen Beschreibungstexte hin, da aufgrund der Randomisierung der Reihenfolge der den VPn angezeigten Aufgaben sowie der Surrogate innerhalb einer Aufgabe ein Reihenfolgeneffekt ausgeschlossen werden kann. Zum anderen können Ursachen für die Unterschiede in der Zusammenstellung der Surrogate einer Aufgabe liegen, wie bereits in Hinblick auf deren thematische Übereinstimmung mit den jeweiligen Beschreibungstexten anerkannt wurde.

Abschließend bleibt als wichtige Erkenntnis zu betonen, dass bei der Entwicklung eines experimentellen Untersuchungsdesigns zur Erforschung von Relevanzkriterien in Hinblick auf das Konstanthalten der thematischen Relevanz bzw. Aboutness als potenzielle Störvariable besondere Vorsicht geboten ist, um eine unerwünschte Konfundierung dieser Drittvariable mit einer höheren Wahrscheinlichkeit ausschließen zu können.

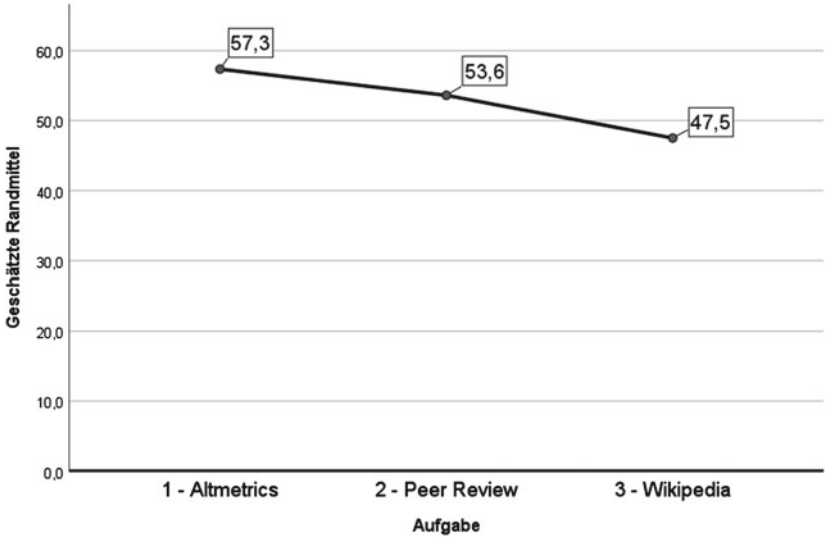

Abbildung 4.25 Mittelwerte der Bewertungen pro Aufgabe

4.6 Grenzen der Studie

Die Grenzen des hier vorgestellten Experiments, in dem mittels Befragung die Daten von Versuchspersonen erhoben wurden, liegen zunächst in den Grenzen quantitativer Erhebungsverfahren allgemein. Die vollstrukturierte Befragung lässt keinen Raum für Flexibilität oder individuell abgestimmte Anpassungen auf die Versuchspersonen; ein solch offenes Vorgehen lässt sich naturgemäß nicht mit explanativen Studien vereinbaren, insbesondere vor dem Hintergrund der Anforderungen an ein echtes Experiment in Hinblick auf die erforderlichen Bedingungen für Kausalität durch Manipulation und Kontrolle.

Die Erhebung der expliziten Relevanzbewertungen in diesem Experiment erfolgte in Form eines multifaktoriellen Online-Survey (Online-Vignettenanalyse), dessen Nachteil darin besteht, dass die zeitliche Präzedenz der Wirkung der UV gegenüber dem beobachteten Effekt der AV als eine der drei Voraussetzungen zum Ableiten kausaler Schlussfolgerungen nicht garantiert werden kann. Der Grund dafür ist, dass mit einer Online-Vignettenanalyse Einstellungen befragt werden bzw. intendiertes anstelle von tatsächlichem Verhalten beobachtet wird (Berger & Wolbring, 2015, S. 46). Es besteht keine Gewissheit

darüber, dass die Versuchspersonen tatsächlich zuerst die Kontextbeschreibungen der Bewertungsaufgaben lasen und im Anschluss die Surrogate bewerteten. Es wäre möglich gewesen, ohne das Lesen der Beschreibungstexte sofort den Bildschirminhalt nach unten zu scrollen. Nichtsdestotrotz besteht die Annahme, dass die Versuchspersonen sich diesbezüglich erwartungsgemäß verhalten haben und die zeitliche Präzedenz als gegeben angesehen werden kann.

Unklar bleibt der Einfluss der Fachdisziplin der Versuchspersonen, da Unterschiede im Publikations- und Zitierverhalten zwischen wissenschaftlichen Forschungskulturen bestehen und die Höhe der Zitationszahlen in verschiedenen Fächern unterschiedlich wertgeschätzt werden könnte. Um den Einfluss unterschiedlicher, fachdisziplinspezifischer Aspekte auf die Relevanzbewertung zu untersuchen, wären diese entsprechend als unabhängige Variablen zu variieren.

Die untersuchte Stichprobe lieferte Daten von 627 Versuchspersonen, die in die Datenanalyse miteinbezogen wurden. Dieser Stichprobenumfang ist im Vergleich zu anderen Studien zu Relevanzkriterien sehr hoch, hauptsächlich aufgrund der designbedingten Einschränkung der Anzahl von Teilnehmenden bei den explorativen Studien, die oft einen qualitativen Ansatz verfolgten. Obgleich des großen Stichprobenumfangs handelt es sich nicht um eine repräsentative Stichprobe, sondern um eine homogene Gruppe in Hinblick auf deren soziodemografische Merkmale (Bildungsstand, Status, Affiliation an einer Universität oder Forschungseinrichtung).

Eine Vielzahl an verschiedenen Merkmalen, Kriterien und Faktoren spielen im Prozess der Relevanzbewertung eine Rolle, die für eine experimentelle Erforschung zwangsläufig auf eine Auswahl bestimmter, als vielversprechend vermuteter Variablen eingegrenzt werden muss. In dem Experiment wurde lediglich eine geringe Anzahl untersucht, d. h. nur eine begrenzte Auswahl an als ursächlich vermutete Variablen wurden in Hinblick auf das Kriterium Popularität für das Stimulusmaterial manipuliert. Dennoch lassen sich die Bewertungen nicht ausschließlich auf diese unabhängigen Variablen zurückführen, da beispielsweise das Publikationsdatum (Kriterium Aktualität) oder Schlüsselwörter im Titel oder Abstract (Kriterium thematische Relevanz) nicht variiert wurden, diese aber dennoch die Bewertung beeinflusst haben können und somit mögliche konfundierende Variablen darstellen, d. h. es ist möglich, dass durch die Vielzahl der verschiedenen Einflussparameter die tatsächlich verwendeten Kriterien bei der Relevanzbewertung nicht aufgedeckt wurden. Die Versuchspersonen haben notwendigerweise in jeder der drei Bewertungsaufgaben jeweils verschiedene Suchergebnisse gesehen, wobei mittels verschiedener Maßnahmen versucht wurde, den Versuchspersonen die Surrogate nach dem Ceteris-Paribus-Prinzip (Döring & Bortz, 2016, S. 99) zur Bewertung vorzulegen: Wie in Abschnitt 4.2.2

ausführlich beschrieben, wurde das Ziel verfolgt, beispielsweise die thematische Relevanz, Aktualität und die Länge des Abstract-Ausschnitts konstant zu halten.

Abschließend sei erwähnt, dass in groß angelegten quantitativen Studien sehr viele Daten erhoben wurden, die naturgemäß im Rahmen einer einzelnen Arbeit nicht alle ausgewertet werden können. Mit der Durchführung der hier vorgestellten Studie liegen neben den experimentell erhobenen Relevanzbewertungen weitere Informationen über die Versuchspersonen vor, die Analysen auf der Ebene der einzelnen Individuen und Einblicke in das akademische Informationssuchverhalten erlauben. Auf derartige tiefergehende, explorative Auswertungen wurde verzichtet, da sie nicht zur Beantwortung der Forschungsfragen der vorliegenden Arbeit beitragen.

Schlussbetrachtungen

5

Die Erforschung von Kriterien, anhand derer informationssuchende Personen die Relevanz von Dokumenten bewerten, ist Gegenstand vieler – hauptsächlich explorativer – informationswissenschaftlicher Studien. In den bisherigen Studien werden unterschiedliche Begriffe als Einflussparameter auf die Relevanzbewertung verwendet, die keiner einheitlichen Definition folgen und sich auch konzeptuell nicht eindeutig voneinander abgrenzen lassen. Diese Forschungslücke wurde im Rahmen der vorliegenden Arbeit mit der Beantwortung der Forschungsfrage F1a (*Wie lassen sich Merkmale, Kriterien und Faktoren als Einflüsse im Prozess der Relevanzbewertung für die Entwicklung eines experimentellen Untersuchungsdesigns definitorisch und konzeptuell voneinander abgrenzen?*) und mit der in direktem Zusammenhang stehenden Forschungsfrage F1b (*Wie können Kriterien bei der Relevanzbewertung von Suchergebnissen für eine experimentelle Studie operationalisiert werden?*) geschlossen. Diese Unterforschungsfragen wurden als Voraussetzung für die Bearbeitung der übergeordneten Forschungsfrage F1 (*Wie können Nutzerkriterien bei der Relevanzbewertung anhand eines experimentellen Untersuchungsdesigns erforscht werden?*) bereits ausführlich im Abschnitt 3.3 erläutert.

In dem nachfolgenden Abschnitt 5.1 erfolgt die Beantwortung der Forschungsfragen F1, F2 und F3. Anschließend werden die Ergebnisse der hier berichteten Forschung im Kontext der Gesamtmethodik reflektiert (Abschnitt 5.2). Die Schlussbetrachtungen schließen mit einem Ausblick auf künftige Forschungsvorhaben ab (Abschnitt 5.3).

C. Behnert, *Popularität und Relevanz in der Suche*,
https://doi.org/10.1007/978-3-658-37512-6_5

5.1 Beantwortung der Forschungsfragen

F1 Wie können Nutzerkriterien bei der Relevanzbewertung anhand eines experimentellen Untersuchungsdesigns erforscht werden?

Die Entwicklung eines experimentellen Untersuchungsdesigns setzt voraus, dass nicht wie mit explorativen Studien offene Forschungsfragen bearbeitet, sondern Hypothesen über operationalisierbare Variablen geprüft werden. Die Nutzerkriterien bei der Relevanzbewertung lassen sich mit den Elementen eines Surrogats (Relevanzmerkmale) operationalisieren, wie beispielsweise das Element Publikationsdatum, anhand dessen das Kriterium Aktualität abgeleitet wird. Relevanzkriterien können also gezielt durch die Variation der (potenziellen) Relevanzmerkmale experimentell untersucht werden, die mithilfe des Modells in Hinblick auf ihre definitorische und konzeptuelle Abgrenzung als Antwort auf die Forschungsfrage F1a (vgl. Abschnitt 3.2) veranschaulicht wurden.

Wie die Ergebnisse der Literaturschau zu den Studien zu Relevanzkriterien zeigen, beeinflussen mehrere Kriterien und somit auch mehrere Elemente im Surrogat in ihrem Zusammenwirken die Relevanzbewertung. Aus diesem Grund ist ein mehrfaktorielles Design einem einfaktoriellen Design vorzuziehen, da nur mit mehrfaktoriellen Designs Interaktionen aufgedeckt und nicht nur die Wirkungen der einzelnen unabhängigen Variablen (Haupteffekte) untersucht werden können. Gerade das hier vorgestellte Experiment verdeutlicht die Notwendigkeit, Interaktionen zu berücksichtigen; wäre die Analyse nur mit Blick auf die Haupteffekte der Anzahl der Downloads (UV 1), der Anzahl der Zitationen des Werks (UV 2) und der Anzahl der Zitationen des Autors (UV 3) auf die Relevanzbewertung als abhängige Variable erfolgt, würden die Ergebnisse zwar ein klares Bild zeigen, jedoch zu falschen Schlussfolgerungen führen. Hinzu kommt, dass erst beim Betrachten der Wechselwirkungen die Ergebnisse der paarweisen Vergleiche die Unterschiede zwischen den statistisch signifikanten und inhaltlich bedeutsamen Differenzwerten aufzeigen; nur bei den Interaktionen zeigen die paarweisen Vergleiche Differenzwerte mit mindestens 10 bzw. 20 Punkten, bei den Einzelwirkungen der unabhängigen Variablen liegen die Differenzwerte maximal bei 6,51 Punkten und sind teilweise sogar vernachlässigbar gering, obwohl sie statistisch signifikant sind (wie zum Beispiel der Wert -1,23 bei $p = 0{,}042$ für UV 2 in Tabelle 4.15).

Eine besondere Herausforderung bei der Planung eines Experiments zur Erforschung von Relevanzkriterien liegt in der Variation des Stimulusmaterials, konkret in der Operationalisierung der einzelnen Stufen der zu untersuchenden unabhängigen Variablen. Das Ziel besteht darin, die Balance zu finden zwischen einer zu

schwachen und einer zu starken Manipulation – erstere kann zu einer zu geringen Effektstärke führen, letztere zu einer verfrühten Offenbarung des eigentlichen Zwecks des Experiments, die ein unerwünschtes Verhalten der Versuchspersonen bewirken und die Ergebnisse in einer Weise beeinflussen kann, die die interne Validität des Experiments gefährdet. Vor diesem Hintergrund ist auch bei der Auswahl der zu untersuchenden Grundgesamtheit zu berücksichtigen, in welchem Kontext diese agiert und über welches Wissen sie vermutlich verfügt.

In dem hier vorgestellten Experiment im akademischen Kontext wurde diese Problematik bei der Auswahl der Surrogate und Suchaufgaben in Abhängigkeit der Probandenakquise erläutert. So wurden gezielt Personen mit einem nicht-informationswissenschaftlichen Hintergrund rekrutiert, die Suchergebnisse zu Themen, mit denen sich die Bibliotheks- und Informationswissenschaft beschäftigt, aus ausschließlich bibliotheks- und informationswissenschaftlichen Quellen bewerten sollen. Auf diese Weise sollte verhindert werden, dass stark manipulierte Daten zu möglicherweise bekannten Autorinnen und Autoren oder Quellen den realen Zweck des Experiments entlarven.

Eine weitere Voraussetzung für die experimentelle Erforschung von Relevanzkriterien wird dahingehend ersichtlich, dass zwischen *predictive judgments* (Bewertungen auf Basis des Surrogats) und *evaluative judgments* (Bewertungen auf Basis des Volltexts) zu unterscheiden ist. Die Operationalisierung der Kriterien anhand der Elemente im Surrogat ist nicht gleichzusetzen mit der Operationalisierung der Kriterien, die bei der Bewertung des eigentlichen Inhalts angewendet werden. In Hinblick auf die in einem experimentellen Design zu manipulierenden unabhängigen Variablen dürfte eine Variation von Volltexten für die Untersuchung des Einflusses bestimmter Aspekte auf *evaluative judgments* die Forschungsleitung vor große Herausforderungen stellen.

Wie oben beschrieben, zeigen die statistischen Ergebnisse der Mehrebenenanalyse ein komplexes Muster, das keine eindeutigen Schlussfolgerungen über den Einfluss der Anzahl der Downloads eines Werks (UV 1), der Zitationszahl eines Werks (UV 2) und der Zitationszahl eines Autors (UV 3) auf die Relevanzbewertung von Surrogaten in akademischen Suchsystemen (AV) zulässt. Demzufolge lassen sich die Forschungsfragen F2 und F3 nicht mit eindeutigen Aussagen beantworten. Nachfolgend werden zum Zweck der Vollständigkeit die beiden Forschungsfragen dennoch getrennt adressiert.

F2 Welchen Einfluss haben Popularitätsdaten auf die Bewertung der Relevanz von Suchergebnissen in akademischen Suchsystemen?

Oberflächlich lässt sich feststellen, dass alle drei UVn einen statistisch signifikanten Effekt auf die Relevanzbewertung ausüben; allerdings besitzen diese Haupteffekte unter Berücksichtigung der diversen Interaktionseffekte keine Gültigkeit. Hieran zeigt sich der Nachteil von Studiendesigns, in denen nur eine einzige unabhängige Variable untersucht oder mehrere unabhängige Variablen nur in Hinblick auf deren Haupteffekte betrachtet werden. Die vermeintlich eindeutigen Ergebnisse führen zu falschen Schlussfolgerungen, wenn nicht zusätzlich das Vorhandensein von Wechselwirkungen überprüft wird. Zudem ergibt sich anhand der paarweisen Vergleiche kein klares Bild über die Richtung eines Effekts. Es lässt sich daher nicht feststellen, ob die untersuchten Popularitätsdaten die Bewertungen der Versuchspersonen positiv oder negativ beeinflusst haben.

F3 Welche Popularitätsdaten beeinflussen die Relevanzbewertung in welchem Maße?

Die Ergebnisse der Literaturschau in Abschnitt 2.1 und die Darstellung der Nutzerkriterien im Prozess der Relevanzbewertung mithilfe des Modells in Abschnitt 3.2 verdeutlichen den besonderen Stellenwert der thematischen Relevanz. Diese stellt die Basis für die Relevanzbewertung dar, auf der weitere Kriterien wie Aktualität und Qualität im Zusammenhang mit Autorität, Glaubwürdigkeit und schließlich Popularität aufsetzen. Der thematischen Relevanz kommt unstrittig die höchste Gewichtung bei der Anwendung von Relevanzkriterien im Prozess der Relevanzbewertung zu. In welcher Weise die weiteren Kriterien gewichtet werden, bleibt unklar. Dass Popularitätsdaten als integrierter Bestandteil der Suchergebnispräsentation in akademischen Suchsystemen als operationalisiertes Kriterium der Popularität gesehen werden können, wurde im Zusammenhang mit der wahrgenommenen Qualität von Suchergebnissen erläutert (vgl. Abschnitt 2.1.2). Welches Maß an Beeinflussung die jeweiligen untersuchten Popularitätsdaten auf die Relevanzbewertung bewirken, kann aufgrund der uneindeutigen statistischen Ergebnisse des Online-Experiments nicht beantwortet werden (vgl. Abschnitt 4.4).

5.2 Reflexion der Ergebnisse im Kontext der Gesamtmethodik

Mit der hier beschriebenen Forschung wurde der Einfluss von Popularitätsdaten als Bestandteil der Suchergebnispräsentation in akademischen Suchsystemen, wie sie in heutigen Systemen wie Google Scholar oder der ACM Digital Library üblich sind, auf die Relevanzbewertung empirisch untersucht. Der Zweck der Untersuchung bestand darin, Kenntnisse über die nutzerseitigen Kriterien, anhand derer informationssuchende Personen die Relevanz von Suchergebnissen bewerten, zu gewinnen. Die Methodik der Arbeit folgte einem quantitativen Forschungsansatz, in dessen Zentrum die Entwicklung und Durchführung eines Online-Experiments stand.

Zunächst wurde mithilfe einer umfassenden Literaturschau zu Studien, in denen Relevanzkriterien erforscht wurden, der aktuelle Stand der Forschung dargelegt. Dabei wurden die bisherigen Studien aus einer inhaltlichen und einer methodischen Perspektive betrachtet und drei konkrete Forschungslücken identifiziert: In den bisherigen Studien wurden (a) die Begriffe Relevanzmerkmale (*relevance clues/cues*), Relevanzkriterien (*relevance criteria*) und Relevanzfaktoren (*relevance factors*) im Kontext der Relevanzbewertung verwendet, die keiner allgemeingültigen Definition folgen und oft nicht klar voneinander abgegrenzt sind, (b) nur selten experimentelle Designs zugrunde gelegt, in denen jedoch Kriterien nicht als konkrete, unabhängige Variablen untersucht wurden, (c) den Teilnehmenden keine Suchergebnisse mit Popularitätsdaten wie die Anzahl von Downloads oder Zitationen eines Werks, wie sie heutzutage Bestandteil moderner akademischer Suchsysteme sind, zur Bewertung vorgelegt.

Diese Lücken sollten mit der Beantwortung der daraus abgeleiteten Forschungsfragen geschlossen werden. Mithilfe der Erkenntnisse aus der Literaturschau wurden die inhaltlichen und methodischen Voraussetzungen für die Entwicklung eines experimentellen Designs zur Erforschung von Relevanzkriterien geschaffen. So erfolgte zunächst die Spezifikation des Relevanzkonzepts, mit der die Definition von Relevanz für diese Arbeit und die Konkretisierung des Prozesses der Relevanzbewertung von Suchergebnissen als ein Prozess des Urteilens einherging. In diesem Zusammenhang wurden die verschiedenen Einflüsse, die im Prozess der Relevanzbewertung eine Rolle spielen, identifiziert.

Zur Veranschaulichung dieser verschiedenen Einflussgrößen wurde ein Modell zur subjektiven Relevanzbewertung von Suchergebnissen in akademischen Suchsystemen erstellt. Hierin liegt ein Neuheitswert der vorliegenden Arbeit, denn mit dem Modell liegt erstmals eine systematische Übersicht über die Elemente von Surrogaten als Relevanzmerkmale, sich daraus abzuleitenden Relevanzkriterien

durch die informationssuchende Person und deren Zusammenwirken sowie die diesen Prozess beeinflussenden Relevanzfaktoren vor. Auf diese Weise trägt das Modell zu einem besseren Verständnis des Bewertungsprozesses bei. Mit der systematischen Darstellung der Surrogatelemente, den subjektiven Kriterien und den Faktoren, die im weiteren Sinn den Kontext der informationssuchenden Person zum Zeitpunkt der Suche bilden, wurde ein Kernproblem der informationswissenschaftlichen Relevanzforschung adressiert. Das Modell stellt somit zugleich eine Definitionsempfehlung der Begriffe Relevanzmerkmale, Relevanzkriterien und Relevanzfaktoren dar und bietet einen theoretischen Rahmen, der für zukünftige Forschungsvorhaben im Kontext des Relevanzbewertungsprozesses berücksichtigt werden sollte. Zusätzlich leistet das Modell einen praktischen Nutzen, indem es als Hilfsmittel zur Operationalisierung in künftigen Untersuchungen zu Relevanzkriterien dienen kann.

Unabhängig von der definitorischen Abgrenzung zwischen Merkmalen, Kriterien und Faktoren berücksichtigt das Modell explizit Popularitätsdaten, wie die Anzahl von Downloads oder Zitationen eines Werkes, die heutzutage in akademischen Suchsystemen wie Google Scholar in die Suchergebnispräsentation integriert sind. Da in bisherigen Studien zu Relevanzkriterien, in denen Jurorinnen und Juroren Surrogate zur Bewertung vorgelegt wurden, solche Popularitätsdaten nicht Bestandteil des Untersuchungsgegenstands waren, wurde im Rahmen dieser Arbeit ein Experiment durchgeführt, um den Einfluss von Popularitätsdaten auf die Relevanzbewertung von Suchergebnissen in akademischen Suchsystemen zu untersuchen. In dieser inhaltlichen Ausrichtung der Studie liegt daher ebenfalls ein Neuheitswert der vorliegenden Arbeit.

Für die Entwicklung des experimentellen Within-Subjects-Designs wurden die Popularitätsdaten operationalisiert als die Anzahl der Downloads (UV 1), die Anzahl der Zitationen des Werks (UV 2) und die Anzahl der Zitationen des Autors (UV 3). Die Datenerhebung erfolgte mithilfe eines Online-Fragebogens, zu dessen Bearbeitung mehr als 700 Forschende an verschiedenen Universitäten in Deutschland motiviert werden konnten. Mithilfe des statistischen Verfahrens der Mehrebenenanalyse wurden in SPSS die Daten von 621 Teilnehmenden ausgewertet. Die Größe der untersuchten Stichprobe übertrifft somit die Stichprobengrößen anderer Studien, in denen explizite Relevanzbewertungen erhoben wurden, um Erkenntnisse über den Prozess der Relevanzbewertung zu gewinnen; sie stellt damit neben der sorgfältigen Zusammenstellung der Stichprobe eine besondere Stärke dieser Arbeit dar.

Die Ergebnisse der statistischen Analyse zeigen, dass es einen statistisch signifikanten Einfluss der Anzahl von Downloads, der Anzahl der Zitationen eines Werks und der Anzahl der Zitationen eines Autors als integrierter Bestandteil

der Suchergebnispräsentation in akademischen Suchsystemen auf die Relevanzbewertung dieser Suchergebnisse gibt. Unklar ist, inwiefern es sich dabei um einen negativen oder positiven Einfluss handelt, d. h. ob die Popularitätsdaten zu einer höheren oder niedrigeren Relevanzbewertung unter sonst gleichen (experimentellen) Bewertungsbedingungen führen.

Rückblickend stellt sich die Frage, ob die Ergebnisse der hier beschriebenen Forschung ein klareres Bild liefern könnten, wenn neben dem Online-Experiment weitere Untersuchungen stattgefunden hätten, also ein Mixed-Methods-Ansatz verfolgt worden wäre.

Mithilfe des Mixed-Methods-Ansatzes werden mehrere Teilstudien unterschiedlicher Designs durchgeführt (Döring & Bortz, 2016, S. 184 f.). Für die hier beschriebene Forschung wäre vermutlich eine qualitative Vorstudie, mit deren Hilfe beispielsweise die Einstellung von Forschenden zu dem Konzept Popularität bzw. zu Popularitätsdaten als Bestandteil von Suchergebnissen explorativ untersucht hätte werden können, sinnvoll gewesen, um weitere Informationen für die Hypothesengenerierung zu erhalten. Für die gezielte Untersuchung des Einflusses von Popularitätsdaten auf die Relevanzbewertung, d. h. zur Beantwortung der Forschungsfragen F2 und F3, wäre ein Mixed-Methods-Ansatz weniger geeignet, da sich nur mit experimentellen, quantitativen Designs ein kausaler Zusammenhang zwischen Ursache und Wirkung durch die systematische Manipulation des Stimulusmaterials und der Kontrolle potenzieller Störvariablen herstellen lässt – an dieser Stelle zeigt sich ein weiterer Neuheitswert der hier beschriebenen Forschung. Allerdings wäre in Hinblick auf die Umsetzung des experimentellen Designs für die Datenerhebung mithilfe des Online-Fragebogens eine zusätzliche Studie zur Evaluierung der ausgewählten Surrogate bezüglich ihrer thematischen Übereinstimmung mit den Beschreibungen der Informationsbedürfnisse nützlich gewesen, um einen inhaltlichen Effekt der Aufgabenbeschreibungen mit höherer Gewissheit ausschließen zu können (vgl. Abschnitt 4.5).

Obwohl die uneindeutigen Befunde des Experiments zunächst den Anschein erwecken mögen, dass die Studie nicht erfolgreich war, leistet sie einen positiven Beitrag für die informationswissenschaftliche Relevanzforschung: Die Ergebnisse liefern einen Hinweis darauf, dass thematische Relevanz auch bei dem Vorhandensein von Popularitätsdaten für die Bewertung von Suchergebnissen das Hauptkriterium ist, anhand dessen die informationssuchende Person die Bewertung vornimmt. Ebenfalls zeigen sie, dass die ausschließliche Untersuchung von Haupteffekten zwar eindeutige, statistisch signifikante Befunde hervorbringen kann, diese jedoch eigentlich keine Aussagekraft besitzen, wenn nicht das Vorhandensein möglicher Wechselwirkungen bereits zu Beginn des Studiendesigns in Betracht gezogen wird.

Ein ebenfalls positiver Beitrag des Experiments liegt in dem empirischen Nachweis über die Komplexität des Relevanzkonzepts und des Relevanzbewertungsprozesses, der sich in dem komplexen Ergebnismuster der statistischen Resultate ausdrückt. Ferner kann das anspruchsvolle und komplexe Design, das im Rahmen dieser Arbeit erstmals zur experimentellen Erforschung von Relevanzkriterien im akademischen Kontext entwickelt wurde, nachgenutzt und adaptiert werden.

5.3 Künftige Forschung

Eine konkrete Erkenntnis für künftige experimentelle Studien zu Relevanzkriterien ergibt sich zunächst in Hinblick auf die Anzahl der zu untersuchenden unabhängigen Variablen (UVn) und der Anzahl ihrer Stufen, auf denen sie variiert werden: Der Vorteil von weniger als drei UVn und jeweils weniger als drei Stufen liegt in einer geringeren Anzahl an experimentellen Bedingungen und daraus resultierend eine geringere Anzahl an Ergebnissen der statistischen paarweisen Vergleiche, wodurch diese Ergebnisse weniger komplex und damit leichter interpretierbar sein können. Eine Möglichkeit wäre, auf eine Stufe wie „keine Angabe“ zu verzichten, wenn die weiteren Stufen eine andere (quantitative) Dimension betreffen, wie es in der hier beschriebenen Studie der Fall ist.

Die Auswertung zeigte wiederum, dass aufgrund der Vielzahl an Merkmalen, Kriterien und Faktoren, welche bei der Relevanzbewertung von Suchergebnissen eine Rolle spielen, ein mehrfaktorielles Design im Vergleich zu mehreren einfaktoriellen Designs in Hinblick auf mögliche Interaktionen vorzuziehen ist. Das bedeutet, dass prinzipiell eher mehr unabhängige Variablen (zu einem Kriterium oder weiteren Kriterien) in einem Experiment variiert werden müssten, damit diese Wechselwirkungen überhaupt aufgedeckt werden können. Unter Umständen erhält die Forschungsleitung dadurch verhältnismäßig viele experimentelle Bedingungen, die, bezogen auf das hier beschriebene Experiment, zu einer hohen Zahl an zu bewertenden Suchergebnissen und zu einem sehr großen zeitlichen Aufwand führen. Probandinnen und Probanden für die Teilnahme an einer zeitlich sehr aufwendigen Studie zu gewinnen und zu einem erfolgreichen Abschluss zu motivieren, dürfte recht schwierig und gegebenenfalls nur mit einem relativ hohen (finanziellen) Anreiz realisierbar sein. Des Weiteren sind unerwünschte Effekte durch Ermüdung zu berücksichtigen.

Die Lösung kann in dem hier gewählten statistischen Auswertungsverfahren liegen: Da die Mehrebenenanalyse fehlende Werte toleriert, kann bereits bei

der Entwicklung des Studiendesigns eingeplant werden, nicht jede Versuchsperson allen Bedingungen im Experiment auszusetzen, sondern gezielt nur einer bestimmten Auswahl an Bedingungen. Dadurch lässt sich der zeitliche Aufwand für die Bearbeitung der Bewertungsaufgaben eingrenzen, ohne dass die Daten an Validität einbüßen.

Der Einfluss der thematischen Relevanz als Fundament der Relevanzbewertung eines Suchergebnisses im akademischen Kontext ist bisher nicht experimentell untersucht worden. Insbesondere in Hinblick auf die Gewichtung weiterer Kriterien, die auf dem Kriterium der thematischen Relevanz aufbauen, ist deren experimentelle Untersuchung vielversprechend. In diesem Zusammenhang sei nochmals auf die in Abschnitt 4.5 diskutierte Notwendigkeit verwiesen, die zu bewertenden Surrogate in Bezug auf die Übereinstimmung ihrer Aboutness als Basis, anhand derer die thematische Relevanz abgeleitet wird, mit den entwickelten Beschreibungstexten zu dem jeweiligen Kontext bzw. Informationsbedürfnis vorab durch Dritte prüfen zu lassen.

Ausgehend von dem Forschungsdesign des hier vorgestellten Experiments im akademischen Kontext ist eine Idee für eine künftige Studie, den Versuchspersonen gezielt Suchergebnisse mit manipulierten Popularitätsdaten, die *nicht* oder *sehr wenig* thematisch relevant sind, zur Bewertung vorzulegen. Die zu prüfende inhaltliche Hypothese könnte lauten: Ist die thematische Relevanz eines Suchergebnisses in Relation zu einem Informationsbedürfnis nicht gegeben, ist die Punktzahl der Relevanzbewertung bei einer hohen Anzahl an Downloads/Werkszitationen/Autorenzitationen höher als bei einer niedrigen Anzahl. Dies würde bedeuten, dass die Basis für die Relevanzbewertung nicht vorhanden wäre. Eine weitere Vermutung wäre, dass das für das Fachgebiet notwendige Wissen zur Ableitung der thematischen Relevanz bei wenig erfahrenen Studierenden (z. B. Erstsemester) fehlt, im Gegensatz zu Promovierenden oder Postdocs. Somit ließen sich zwei Gruppen mit einem unterschiedlichen Erfahrungsstand miteinander vergleichen.

Ein weiteres Experiment zur Untersuchung der Abhängigkeit von Bewertungen von thematischer Relevanz auch unabhängig von einem akademischen Kontext könnte als Between-Subjects-Design konzipiert werden, in welchem drei Gruppen verschiedenartige Suchergebnisse zur Bewertung vorgelegt werden: eine Experimentalgruppe erhält Suchergebnisse mit Autorennamen und Erscheinungsjahr und ausgewählten Popularitätsdaten, aber ohne Abstract und eventuell ohne Titel, eine zweite Experimentalgruppe bekommt die gleichen Suchergebnisse mit Titel, Abstract und Popularitätsdaten, einer Kontrollgruppe werden die gleichen Suchergebnisse mit Titel und Abstract, jedoch ohne Popularitätsdaten präsentiert.

Die Annahme ist, dass nur die Kontrollgruppe überhaupt in der Lage ist, die thematische Relevanz beurteilen zu können; diesbezüglich können die Unterschiede in der Punktzahl der Relevanzbewertungen aus den drei Teilstichproben zeigen, wie nah Relevanzbewertungen der Suchergebnisse ohne Popularitätsdaten im Vergleich zu denen mit Popularitätsdaten an den als thematisch relevant beurteilten Suchergebnissen liegen.

Schließlich können künftige Studien die im Rahmen dieser Arbeit erhobenen Rohdaten[1] nachnutzen, zum Beispiel für:

- Explorative Analysen, um mögliche Korrelationen aufzudecken;
- Die gezielte Analyse der mithilfe des Vorab- und Anschlussfragebogens erhobenen Daten bezüglich des Informationssuchverhaltens von Forschenden, beispielsweise in Hinblick auf die Nutzung von akademischen Suchsystemen oder die Einstellung gegenüber den Kriterien bei der Relevanzbewertung insbesondere in Bezug auf die in der Arbeit untersuchten Popularitätsdaten;
- Studien zu Replikationszwecken.

Die weitere Untersuchung des Einflusses von Popularitätsdaten ist im Kontext der akademischen Informationssuche von besonderem Interesse, weil anhand solcher potenzieller Relevanzmerkmale das Kriterium der Popularität und wiederum das der Qualität abgeleitet werden können. Insbesondere im akademischen Kontext ist davon auszugehen, dass Informationsobjekte neben der thematischen Relevanz vordergründig nach ihrer (vermuteten) Qualität beurteilt werden. Thematische Relevanz ist von einer informationssuchenden Person jedoch mitunter schwierig zu beurteilen, wenn sie nicht über das notwendige Wissen über das Thema verfügt. Im akademischen Kontext betrifft dies zum Beispiel eher unerfahrene Forschende, die am Beginn ihrer Promotionsforschung stehen, und nicht so sehr Forschende, die bereits über mehrere Jahre auf einem wissenschaftlichen Gebiet Erfahrungen gesammelt haben. Wenn thematische Relevanz als das Basiskriterium der Relevanzbewertung nicht in ausreichender Güte bedient werden kann, stellt sich die Frage, an welchen Merkmalen und Kriterien sich informationssuchende Personen stattdessen orientieren, um die Qualität eines Werkes anhand seines Surrogats abzuleiten.

Allerdings können Popularitätsdaten wie Angaben zu Zitations- und Downloadhäufigkeiten einen Matthäuseffekt herbeiführen, d. h. bereits vielzitierte Werke werden als qualitativ wertvoller erachtet und erlangen weitere Zitationen, während wenig zitierten Werken eine geringere Qualität zugesprochen wird, was

[1] Siehe https://doi.org/10.17605/OSF.IO/NTWQD.

wiederum ein Grund für ausbleibende Zitationen sein kann. In Bezug auf die Anzahl an Downloads bedeutet dies, dass eine hohe Anzahl zu weiteren Downloads führt und die Zahl der Downloads sich weiter erhöht. Eine Beurteilung der (vermuteten) Qualität eines Werkes anhand solcher Popularitätsdaten entspricht nicht den vier Kriterien für Qualität in der Wissenschaft (inhaltliche Relevanz, methodische Strenge, ethische Strenge, Präsentationsqualität) (Döring & Bortz, 2016, S. 90), welche weniger auf Basis von Surrogaten, sondern ausschließlich anhand des Volltexts vollständig beurteilt werden können und ein hohes Maß an Wissenschaftskompetenz und Erfahrung erfordern. Umso wichtiger ist die weitere Erforschung von Relevanzkriterien bei der Bewertung von Suchergebnissen, wenn die Beurteilung von Qualität maßgeblich anhand der Popularität – über einen akademischen Kontext hinausgehend – nicht als erstrebenswert gilt.

Neben der Erforschung von Relevanzkriterien sollte in künftigen Studien der Einfluss von Relevanzfaktoren in Kombination mit ausgewählten Kriterien untersucht werden. Insbesondere der disziplinspezifische Hintergrund der Versuchspersonen ist – wie in Abschnitt 4.6 beschrieben – in Hinblick auf die Beurteilung von Zitationszahlen von besonderem Interesse, ebenso wie ihr soziokultureller Kontext bezüglich des Einflusses von Machtdistanz bzw. Autoritätswahrnehmung auf die Anwendung der Kriterien Popularität und kognitive Autorität. Anregungen für Relevanzmerkmale und Relevanzkriterien als potenzielle unabhängige Variablen kann hier das in Abschnitt 3.2 vorgestellte Modell zur subjektiven Relevanzbewertung von Suchergebnissen in akademischen Suchsystemen liefern. Generell ist zu bedenken: Mit jedem neuen Element, das künftig in die Suchergebnisdarstellung von (akademischen) Suchsystemen integriert wird, liegt ein weiteres potenzielles Relevanzmerkmal vor, dessen Effekt auf die Relevanzbewertung zu untersuchen wäre.

Literatur- und Quellenverzeichnis

Agichtein, E., Brill, E., & Dumais, S. (2006). Improving web search ranking by incorporating user behavior information. *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval – SIGIR '06*, 19–26. https://doi.org/10.1145/1148170.1148177

Albassam, S. A. A., & Ruthven, I. (2018). Users' relevance criteria for video in leisure contexts. *Journal of Documentation*, *74*(1), 62–79. https://doi.org/10.1108/JD-06-2017-0081

Asher, A. D., Duke, L. M., & Wilson, S. (2013). Paths of discovery: Comparing the search effectiveness of EBSCO Discovery Service, Summon, Google Scholar, and conventional library resources. *College & Research Libraries*, *74*(5), 464–488.

Auspurg, K., & Hinz, T. (2007). Multifactorial experiments in surveys: Conjoint Analysis, Choice Experiments, and Factorial Surveys. In M. Keuschnigg & T. Wolbring (Hrsg.), *Experimente in den Sozialwissenschaften: Bd. Sonderband* (S. 291–315). Nomos.

Baeza-Yates, R., & Ribeiro-Neto, B. (2011). *Modern information retrieval : the concepts and technology behind search* (2. ed.). Addison-Wesley/Pearson.

Bailey, E. (2017). *Measuring online search expertise* [University of North Carolina at Chapel Hill]. https://cdr.lib.unc.edu/record/uuid:6086308c-0965-458f-a593-a542fc0a2f9c

Bailey, E., & Kelly, D. (2016). Developing a Measure of Search Expertise. *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval – CHIIR '16*, 237–240. https://doi.org/10.1145/2854946.2854983

Balatsoukas, P., & Demian, P. (2009). Effects of granularity of search results on the relevance judgment behavior of engineers: Building systems for retrieval and understanding of context. *Journal of the American Society for Information Science and Technology*, *14*(4), 453–467. https://doi.org/10.1002/asi.21268

Balatsoukas, P., O'Brien, A., & Morris, A. (2010). Design factors affecting relevance judgment behaviour in the context of metadata surrogates. *Journal of Information Science*, *36*(6), 780–797. https://doi.org/10.1177/0165551510386174

Balatsoukas, P., & Ruthven, I. (2012). An eye-tracking approach to the analysis of relevance judgments on the Web: The case of Google search engine. *Journal of the American Society for Information Science and Technology*, *63*(9), 1728–1746. https://doi.org/10.1002/asi.22707

C. Behnert, *Popularität und Relevanz in der Suche*,
https://doi.org/10.1007/978-3-658-37512-6

Balatsoukas, P., & Ruthven, I. (2010). The use of relevance criteria during predictive judgment: An eye tracking approach. *Proceedings of the American Society for Information Science and Technology*, *47*, 1–10. https://doi.org/10.1002/meet.14504701145

Bales, S., & Wang, P. (2006). Consolidating user relevance criteria: A meta-ethnography of empirical studies. *Proceedings of the American Society for Information Science and Technology*, *42*. https://doi.org/10.1002/meet.14504201277

Baltes-Götz, B. (2019). *Analyse von hierarchischen linearen Modellen mit SPSS* (Bd. 2019). Zentrum für Informations-, Medien- und Kommunikationstechnologie (ZIMK) an der Universität Trier. https://www.uni-trier.de/fileadmin/urt/doku/hlm/hlm.pdf

Barry, C. L. (1994). User-defined relevance criteria: An exploratory study. *Journal of the American Society for Information Science*, *45*(3), 149–159. https://doi.org/10.1002/(SICI)1097-4571(199404)45:3<149::AID-ASI5>3.0.CO;2-J

Barry, C. L. (1998). Document representations and clues to document relevance. *Journal of the American Society for Information Science*, *49*(14), 1293–1303. https://doi.org/10.1002/(SICI)1097-4571(1998)49:14<1293::AID-ASI7>3.0.CO;2-E

Barry, C. L., & Schamber, L. (1998). Users' criteria for relevance evaluation: A cross-situational comparison. *Information Processing & Management*, *34*(2–3), 219–236. https://doi.org/10.1016/S0306-4573(97)00078-2

Barry, C., & Lardner, M. (2011). A study of first click behaviour and user interaction on the Google SERP. In J. Pokorny, V. Repa, K. Richta, W. Wojtkowski, H. Linger, C. Barry, & M. Lang (Hrsg.), *Information Systems Development* (S. 89–99). Springer New York. https://doi.org/10.1007/978-1-4419-9790-6_7

Behnert, C. (2019). Kriterien und Einflussfaktoren bei der Relevanzbewertung von Surrogaten in akademischen Informationssystemen. *Information – Wissenschaft & Praxis*, *70*(1), 24–32. https://doi.org/10.1515/iwp-2019-0002

Behnert, C., & Lewandowski, D. (2015). Ranking search results in library information systems — Considering ranking approaches adapted from web search engines. *The Journal of Academic Librarianship*, *41*(6), 725–735. https://doi.org/10.1016/j.acalib.2015.07.010

Belkin, N. J. (1980). Anomalous states of knowledge as a basis for information retrieval. *Canadian Journal of Information Science*, *5*, 133–143.

Belkin, N. J. (2015). People, interacting with information. *ACM SIGIR Forum*, *49*(2), 13–27. https://doi.org/10.1145/2888422.2888424

Beresi, U. C. (2011). *Related scientific information: A study on user-defined relevance* [Robert Gordon University]. http://hdl.handle.net/10059/705

Beresi, U. C., Kim, Y., Song, D., Ruthven, I., & Baillie, M. (2010). Relevance in Technicolor. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Bd. 6273 LNCS* (S. 196–207). https://doi.org/10.1007/978-3-642-15464-5_21

Berger, R., & Wolbring, T. (2015). Kontrafaktische Kausalität und eine Typologie sozialwissenschaftlicher Experimente. In M. Keuschnigg & T. Wolbring (Hrsg.), *Experimente in den Sozialwissenschaften* (S. 34–52). Nomos.

Bookstein, A. (1979). Relevance. *Journal of the American Society for Information Science*, *30*(5), 269–273. https://doi.org/https://doi.org/10.1002/asi.4630300505

Borlund, P. (2003a). The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. *Information Research*, *8*(3). http://www.informationr.net/ir/8-3/paper152.html

Borlund, P. (2003b). The concept of relevance in IR. *Journal of the American Society for Information Science and Technology*, *54*(10), 913–925. https://doi.org/10.1002/asi.10286

Borlund, P. (2013). Interactive Information Retrieval: An Introduction. *Journal of Information Science Theory and Practice*, *1*(3), 12–32. https://doi.org/10.1633/JISTaP.2013.1.3.2

Borlund, P. (2016). A study of the use of simulated work task situations in interactive information retrieval evaluations. *Journal of Documentation*, *72*(3), 394–413. https://doi.org/10.1108/JD-06-2015-0068

Borlund, P., & Ingwersen, P. (1997). The development of a method for the evaluation of interactive information retrieval systems. *Journal of Documentation*, *53*(3), 225–250. https://doi.org/10.1108/EUM0000000007198

Bortz, J., & Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler* (7., vollst). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-12770-0

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, *30*(1–7), 107–117. https://doi.org/10.1016/S0169-7552(98)00110-X

Broder, A. (2002). A taxonomy of web search. *ACM SIGIR Forum*, *36*(2), 3–10. https://doi.org/10.1145/792550.792552

Bruce, H. W. (1994). A cognitive view of the situational dynamism of user-centered relevance estimation. *Journal of the American Society for Information Science*, *45*(3), 142–148. https://doi.org/10.1002/(SICI)1097-4571(199404)45:3<142::AID-ASI4>3.0.CO;2-6

Brunswik, E. (1952). *The conceptual framework of psychology*. University of Chicago Press.

Bruza, P., & Chang, V. (2014). Perceptions of document relevance. *Frontiers in Psychology*, *5*(JUL), 1–8. https://doi.org/10.3389/fpsyg.2014.00612

Bruza, P. D., Song, D. W., & Wong, K. F. (2000). Aboutness from a commonsense perspective. *Journal of the American Society for Information Science and Technology*, *51*(12), 1090–1105. https://doi.org/10.1002/1097-4571(2000)9999:9999<::AID-ASI1026>3.0.CO;2-Y

Buckland, M. K. (1991). Information as thing. *Journal of the American Society for Information Science*, *42*(5), 351–360. https://doi.org/10.1002/(SICI)1097-4571(199106)42:5<351::AID-ASI5>3.0.CO;2-3

Buckland, M. K. (2017). *Information and society*. MIT Press.

Buckley, C., & Voorhees, E. M. (2005). Retrieval system evaluation. In E. M. Voorhees & D. K. Harman (Hrsg.), *TREC: Experiment and Evaluation in Information Retrieval* (S. 53–75). MIT Press.

Byström, K., & Hansen, P. (2005). Conceptual framework for tasks in information studies. *Journal of the American Society for Information Science and Technology*, *56*(10), 1050–1061. https://doi.org/10.1002/asi.20197

Case, D. O., & Given, L. M. (2016). *Looking for information: a survey of research on information seeking, needs, and behavior* (4. ed.). Emerald.

Choi, Y., & Rasmussen, E. M. (2002). Users' relevance criteria in image retrieval in American history. *Information Processing and Management*, *38*(5), 695–726. https://doi.org/10.1016/S0306-4573(01)00059-0

Clemmensen, M. L., & Borlund, P. (2016). Order effect in interactive information retrieval evaluation: an empirical study. *Journal of Documentation*, *72*(2), 194–213. https://doi.org/10.1108/JD-04-2015-0051

Cole, M., Liu, J., Belkin, N. J., Bierig, R., Gwizdka, J., Liu, C., Zhang, J., & Zhang, X. (2009). Usefulness as the criterion for evaluation of interactive information retrieval systems. *Proceedings of the Third Workshop on Human-Computer Interaction and Information Retrieval (HCIR 2009)*, 1–4. http://cuaslis.org/hcir2009/HCIR2009.pdf

Connaway, L. S., & Radford, M. L. (2017). *Research methods in library and information science* (6.ed.). Libraries Unlimited.

Cook, K. H. (1971). *A predictive model of human relevance decisions.* Syracuse University; Department of Mass Communications.

Cool, C., & Belkin, N. J. (2011). Interactive information retrieval: history and background. In I. Ruthven & D. Kelly (Hrsg.), *Interactive information seeking, behaviour and retrieval* (S. 1–14). Facet.

Cool, C., Belkin, N. J., Kantor, P. B., & Frieder, O. (1993). Characteristics of texts affecting relevance judgments. *Proceedings of the 14th Annual National Online Meeting, New York, May 4 – 6, 1993*, 77–84.

Cooper, W. S. (1971). A definition of relevance for information retrieval. *Information Storage and Retrieval*, *7*(1), 19–37. https://doi.org/10.1016/0020-0271(71)90024-6

Cosijn, E. (2009). Relevance Judgments and Measurements. In *Encyclopedia of Library and Information Sciences, Third Edition* (Nummer August 2015, S. 4512–4519). CRC Press. https://doi.org/10.1081/E-ELIS3-120044537

Cosijn, E., & Ingwersen, P. (2000). Dimensions of relevance. *Information Processing & Management*, *36*(4), 533–550. https://doi.org/10.1016/S0306-4573(99)00072-2

Crescenzi, A., Kelly, D., & Azzopardi, L. (2016). Impacts of time constraints and system delays on user experience. *CHIIR 2016 – Proceedings of the 2016 ACM Conference on Human Information Interaction and Retrieval*, 141–150. https://doi.org/10.1145/2854946.2854976

Crystal, A., & Greenberg, J. (2006). Relevance criteria identified by health information users during Web searches. *Journal of the American Society for Information Science and Technology*, *57*(10), 1368–1382. https://doi.org/10.1002/asi.20436

Cuadra, C. A., & Katter, R. V. (1967a). Opening the black box of 'relevance'. *Journal of Documentation*, *23*(4), 291–303. https://doi.org/10.1108/eb026436

Cuadra, C. A., & Katter, R. V. (1967b). The relevance of relevance assessment. *Proceedings of the American Documentation Institute : annual meeting*, 95–99.

Dervin, B., & Nilan, M. (1986). Information needs and uses. *Annual Review of Information Science and Technology*, *21*, 3–33. https://doi.org/10.1016/S0022-5371(63)80069-9

Döring, N. (2013). Zur Operationalisierung von Geschlecht im Fragebogen: Probleme und Lösungsansätze aus Sicht von Mess-, Umfrage-, Gender- und Queer-Theorie. *Gender*, *2*(2), 94–113. https://doi.org/10.1002/9783527662852.ch1

Döring, N., & Bortz, J. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften* (5. Aufl.). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-41089-5

Eid, M., Gollwitzer, M., & Schmitt, M. (2017). *Statistik und Forschungsmethoden* (5., korr.). Beltz.

Eisenberg, M. B. (1988). Measuring relevance judgments. *Information Processing and Management*, *24*(4), 373–389. https://doi.org/10.1016/0306-4573(88)90042-8

Eisenberg, M. B., & Hu, X. (1987). Dichotomous relevance judgments and the evaluation of information systems. *Proceedings of the American Society for Information Science*, 66–69.

Eisenberg, M., & Barry, C. (1988). Order effects: A study of the possible influence of presentation order on user judgments of document relevance. *Journal of the American Society for Information Science*, *39*(5), 293–300. https://doi.org/10.1002/(SICI)1097-4571(198809)39:5<293::AID-ASI1>3.0.CO;2-I

Ellis, D. (1989). A behavioural approach to information retrieval system design. *Journal of Documentation*, *45*(3), 171–212. https://doi.org/10.1108/eb026843

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Fitzgerald, M. A., & Galloway, C. (2001). Relevance judging, evaluation, and decision making in virtual libraries: A descriptive study. *Journal of the American Society for Information Science and Technology*, *52*(12), 989–1010. https://doi.org/10.1002/asi.1152

Freund, L., & Wildemuth, B. M. (2014). Documenting and studying the use of assigned search tasks: RepAST. *Proceedings of the ASIST Annual Meeting*, *51*(1). https://doi.org/10.1002/meet.2014.14505101122

Froehlich, T. J. (1994). Relevance reconsidered – towards an agenda for the 21st century: Introduction to special topic issue on relevance research. *Journal of the American Society for Information Science*, *45*(3), 124–134.

Funke, F. (2016). A Web Experiment Showing Negative Effects of Slider Scales Compared to Visual Analogue Scales and Radio Button Scales. *Social Science Computer Review*, *34*(2), 244–254. https://doi.org/10.1177/0894439315575477

Funke, F., Reips, U. D., & Thomas, R. K. (2011). Sliders for the smart: Type of rating scale on the web interacts with Educational Level. *Social Science Computer Review*, *29*(2), 221–231. https://doi.org/10.1177/0894439310376896

Galton, F. (1907). Vox populi. *Nature*, *75*(1949), 450–451. https://www.nature.com/articles/075450a0.pdf

Greisdorf, H. (2000). Relevance: An interdisciplinary and information science perspective. *Informing Science*, *3*(2), 67–71.

Greisdorf, H. (2003). Relevance thresholds: A multi-stage predictive model of how users evaluate information. *Information Processing & Management*, *39*(3), 403–423. https://doi.org/10.1016/S0306-4573(02)00032-8

Haider, J., & Sundin, O. (2019). *Invisible search and online search engines: The ubiquity of search in everyday life*. Routledge.

Hamid, R. A., Thom, J. A., & Iskandar, D. A. (2016). Effects of relevance criteria and subjective factors on web image searching behaviour. *Journal of Information Science*, 1–15. https://doi.org/10.1177/0165551516666968

Harman, D. K. (2005). The TREC Test Collections. In E. M. Voorhees & D. K. Harman (Hrsg.), *TREC: Experiment and Evaluation in Information Retrieval* (S. 21–52). MIT Press. https://doi.org/10.1162/coli.2006.32.4.563

Harter, S. P. (1992). Psychological relevance and information science. *Journal of the American Society for Information Science*, *43*(9), 602–615. https://doi.org/10.1002/(SICI)1097-4571(199210)43:9<602::AID-ASI3>3.0.CO;2-Q

Haustein, S. (2012). *Multidimensional journal evaluation: Analyzing scientific periodicals beyond the Impact Factor*. De Gruyter Saur.

Heinström, J. (2003). Five personality dimensions and their influence on information behaviour. *Information Research-an International Electronic Journal*, *9*(1), 165. http://www.informationr.net/ir/9-1/paper165.html

Heinström, J. (2005). Fast surfing, broad scanning and deep diving: The influence of personality and study approach on students' information-seeking behavior. *Journal of Documentation*, *61*(2), 228–247. https://doi.org/10.1108/00220410510585205

Hirsh, S. G. (1999). Children's relevance criteria and information seeking on electronic resources. *Journal of the American Society for Information Science*, *50*(14), 1265–1283. https://doi.org/10.1002/(SICI)1097-4571(1999)50:14<1265::AID-ASI2>3.0.CO;2-E

Hjørland, B. (2000). Relevance research: The missing perspective(s): „Non-relevance" and „epistemological relevance". *Journal of the American Society for Information Science*, *51*(2), 209–211. https://doi.org/10.1002/(SICI)1097-4571(2000)51:2<209::AID-ASI14>3.0.CO;2-B

Hjørland, B. (2001). Towards a theory of aboutness, subject, topicality, theme, domain, field, content . . . and relevance. *Journal of the American Society for Information Science and Technology*, *52*(9), 774–778. https://doi.org/10.1002/asi.1131

Hjørland, B. (2010). The foundation of the concept of relevance. *Journal of the American Society for Information Science and Technology*, *61*(2), 217–237. https://doi.org/10.1002/asi.21261

Hjørland, B., & Christensen, F. S. (2002). Work tasks and socio-cognitive relevance: A specific example. *Journal of the American Society for Information Science and Technology*, *53*(11), 960–965. https://doi.org/10.1002/asi.10132

Hoffman, L., & Rovine, M. J. (2007). Multilevel models for the experimental psychologist: Foundations and illustrative examples. *Behavior Research Methods*, *39*(1), 101–117. https://doi.org/10.3758/BF03192848

Hofstede, G., Hofstede, G. J., & Minkov, M. (2017). *Lokales Denken, globales Handeln: Interkulturelle Zusammenarbeit und globales Management* (6. Aufl.). dtv.

Hogarth, R. M. (1987). *Judgment and choice: The psychology of decision* (2nd Ed.). Wiley.

Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology*, *24*(1), 1–55. https://doi.org/10.1016/0010-0285(92)90002-J

Howard, D. L. (1994). Pertinence as reflected in personal constructs. *Journal of the American Society for Information Science*, *45*(3), 172–185. https://doi.org/10.1002/(SICI)1097-4571(199404)45:3<172::AID-ASI7>3.0.CO;2-V

Huang, M., & Wang, H. (2004). The influence of document presentation order and number of documents judged on users' judgments of relevance. *Journal of the American Society for Information Science and Technology*, *55*(11), 970–979. https://doi.org/10.1002/asi.20047

Huang, X., & Soergel, D. (2013). Relevance: An improved framework for explicating the notion. *Journal of the American Society for Information Science and Technology*, *64*(1), 18–35. https://doi.org/10.1002/asi.22811

Jacsó, P. (2005). Google Scholar: the pros and the cons. *Online Information Review*, *29*(2), 208–214. https://doi.org/10.1108/14684520510598066

Janes, J. W. (1991a). Relevance judgments and the incremental presentation of document representations. *Information Processing and Management*, *27*(6), 629–646. https://doi.org/10.1016/0306-4573(91)90004-6

Janes, J. W. (1991b). The binary nature of continuous relevance judgments: A study of users' perceptions. *Journal of the American Society for Information Science*, *42*(10), 754–756. https://doi.org/10.1002/(SICI)1097-4571(199112)42:10<754::AID-ASI9>3.0.CO;2-C

Jansen, B. J. B. J., & Spink, A. (2006). How are we searching the World Wide Web? A comparison of nine search engine transaction logs. *Information Processing & Management*, *42*(1), 248–263. https://doi.org/10.1016/j.ipm.2004.10.007

Joachims, T., Granka, L., Pan, B., Hembrooke, H., & Gay, G. (2005). Accurately interpreting clickthrough data as implicit feedback. *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval – SIGIR '05*, 154–161. https://doi.org/10.1145/1076034.1076063

Kammerer, Y., & Gerjets, P. (2014). The role of search result position and source trustworthiness in the selection of web search results when using a list or a grid interface. *International Journal of Human-Computer Interaction*, *30*(3), 177–191. https://doi.org/10.1080/10447318.2013.846790

Katter, R. V. (1968). The influence of scale form on relevance judgments. *Information Storage and Retrieval*, *4*(1), 1–11. https://doi.org/10.1016/0020-0271(68)90002-8

Kekäläinen, J., & Järvelin, K. (2002). Using graded relevance assessments in IR evaluation. *Journal of the American Society for Information Science and Technology*, *53*(13), 1120–1129. https://doi.org/10.1002/asi.10137

Kelly, D. (2009). Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends® in Information Retrieval*, *3*(1—2). https://doi.org/10.1561/1500000012

Kelly, D., & Belkin, N. J. (2004). Display time as implicit feedback. *Proceedings of the 27th annual international conference on Research and development in information retrieval – SIGIR '04*, 377–384. https://doi.org/10.1145/1008992.1009057

Kelly, D., & Cresenzi, A. (2016). From design to analysis: Conducting controlled laboratory experiments with users. *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval – SIGIR '16*, 1207–1210. https://doi.org/10.1145/2911451.2914809

Kemp, D. A. (1974). Relevance, pertinence and information system development. *Information Storage and Retrieval*, *10*(2), 37–47. https://doi.org/10.1016/0020-0271(74)90002-3

Königová, M. (1971). Mathematical and statistical methods of noise evaluation in a retrieval system. *Information Storage and Retrieval*, *6*(6), 437–444. https://doi.org/10.1016/0020-0271(71)90009-X

Kurtz, M. J., & Bollen, J. (2010). Usage bibliometrics. *Annual Review of Information Science and Technology*, *44*(1), 1–64. https://doi.org/10.1002/aris.2010.1440440108

Lenhard, W., & Lenhard, A. (2016). *Berechnung von Effektstärken.* Psychometrica. https://doi.org/10.13140/RG.2.1.3478.4245

Lewandowski, D. (2008). The retrieval effectiveness of web search engines: considering results descriptions. *Journal of Documentation*, *64*(6), 915–937. https://doi.org/10.1108/00220410810912451

Lewandowski, D. (2009). Ranking library materials. *Library Hi Tech*, *27*(4), 584–593. https://doi.org/10.1108/07378830911007682

Lewandowski, D. (2010). Der OPAC als Suchmaschine. In J. Bergmann & P. Danowski (Hrsg.), *Handbuch Bibliothek 2.0* (S. 87–107). De Gruyter Saur.

Lewandowski, D. (2012). Credibility in web search engines. In S. Apostel & M. Folk (Hrsg.), *Online Credibility and Digital Ethos: Evaluating Computer – Mediated Communication* (S. 131–145). IGI Global.

Maddalena, E., Mizzaro, S., Scholer, F., & Turpin, A. (2017). On Crowdsourcing Relevance Magnitudes for Information Retrieval Evaluation. In *ACM Transactions on Information Systems* (Bd. 35, Nummer 3). https://doi.org/10.1145/3002172

Maglaughlin, K. L., & Sonnenwald, D. H. (2002). User perspectives on relevance criteria: A comparison among relevant, partially relevant, and not-relevant judgments. *Journal of the American Society for Information Science and Technology*, *53*(5), 327–342. https://doi.org/10.1002/asi.10049

Mansourian, Y., & Ford, N. (2007). Search persistence and failure on the web: a "bounded rationality" and "satisficing" analysis. *Journal of Documentation*, *63*(5), 680–701. https://doi.org/10.1108/00220410710827754

Maron, M. E. (1977). On indexing, retrieval and the meaning of about. *Journal of the American Society for Information Science*, *28*(1), 38–43. https://doi.org/10.1002/asi.4630280107

Merton, R. K. (1968). The Matthew Effect in Science: The reward and communication systems of science are considered. *Science*, *159*(3810), 56–63. https://doi.org/10.1126/science.159.3810.56

Mizzaro, S. (1997). Relevance: The whole history. *Journal of the American Society for Information Science*, *48*(9), 810–832. https://doi.org/10.1002/(SICI)1097-4571(199709)48:9<810::AID-ASI6>3.0.CO;2-U

Montgomery, H. (1983). Decision Rules and the Search for a Dominance Structure: Towards a Process Model of Decision Making. *Advances in Psychology*, *14*, 343–369. https://doi.org/10.1016/S0166-4115(08)62243-8

Nezlek, J. B., Schröder-Abé, M., & Schütz, A. (2006). Mehrebenenanalysen in der psychologischen Forschung: Vorteile und Möglichkeiten der Mehrebenenmodellierung mit Zufallskoeffizienten. *Psychologische Rundschau*, *57*(4), 213–223. https://doi.org/10.1026/0033-3042.57.4.213

Nicholas, D., Huntington, P., Jamali, H. R., & Dobrowolski, T. (2008). The information-seeking behaviour of the digital consumer: case study–the virtual scholar. In D. Nicholas & I. Rowlands (Hrsg.), *Digital Consumers: reshaping the information profession* (S. 113–158). Facet.

O'Neill, E. T., Kammerer, K. A., & Bennett, R. (2017). The aboutness of words. *Journal of the Association for Information Science and Technology*, *68*(1992), 2471–2483. https://doi.org/10.1002/asi.23856

OECD. (2007). Revised fields of science and technology (FOS) in the Frascati Manual. In *Working Party of National Experts on Science and Technology Indicators (NETSI)* (S. 12).

Organisation for Economic Co-operation and Development (OECD). http://www.oecd.org/science/inno/38235147.pdf

Olaisen, J. (1990). Information quality factors and the cognitive authority of electronic information. In I. Wormell (Hrsg.), *Information quality: Definitions and dimensions; Proceedings of a NORDINFO seminar, Royal School of Librarianship, Copenhagen, 1989* (S. 91–121). Taylor Graham.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1998). The PageRank citation ranking: bringing order to the web. In *Technical report, Stanford Digital Library Technologies Project*. Stanford University. http://ilpubs.stanford.edu:8090/422/

Pan, B., Hembrooke, H., Joachims, T., Lorigo, L., Gay, G., & Granka, L. (2007). In Google we trust: Users' decisions on rank, position, and relevance. *Journal of Computer-Mediated Communication*, *12*(3), 801–823. https://doi.org/10.1111/j.1083-6101.2007.00351.x

Petty, R. E., Briñol, P., Loersch, C., & McCaslin, M. J. (2014). The Need for Cognition. In M. R. Leary & R. H. Hoyle (Hrsg.), *Handbook of Individual Differences in Social Behavior* (S. 116–131). Guilford Publications.

Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. *Advances in experimental social psychology*, *19*, 123–205.

Pirolli, P., & Card, S. (1999). Information foraging. *Psychological Review*, *106*(4), 643–675. https://doi.org/10.1037/0033-295X.106.4.643

Plassmeier, K., Borst, T., Behnert, C., & Lewandowski, D. (2015). Evaluating popularity data for relevance ranking in library information systems. *Proceedings of the Association for Information Science and Technology*, *52*(1), 1–4. https://doi.org/10.1002/pra2.2015.145052010012 5

Plessner, H. (2017a). Entscheiden, Entscheidungstheorie. In *Dorsch – Lexikon der Psychologie* (18., übera, S. 484). Hogrefe.

Plessner, H. (2017b). Urteilen. In *Dorsch – Lexikon der Psychologie* (18., übera, S. 1757). Hogrefe.

Prabha, C., Silipigni Connaway, L., Olszewski, L., & Jenkins, L. R. (2007). What is enough? Satisficing information needs. *Journal of Documentation*, *63*(1), 74–89. https://doi.org/10.1108/00220410710723894

Purgailis Parker, L. M., & Johnson, R. E. (1990). Does order of presentation affect users' judgment of documents? *Journal of the American Society for Information Science*, *41*(7), 493–494. https://doi.org/10.1002/(SICI)1097-4571(199010)41:7<493::AID-ASI2>3.0.CO;2-0

Regazzi, J. J. (1988). Performance measures for information retrieval systems—an experimental approach. *Journal of the American Society for Information Science*, *39*(4), 235–251. https://doi.org/10.1002/(SICI)1097-4571(198807)39:4<235::AID-ASI3>3.0.CO;2-H

Reid, J. (1999). A new, task-oriented paradigm for information retrieval: Implications for evaluation of information retrieval systems. In T. Aparac, T. Saracevic, P. Ingwersen, & P. Vakkari (Hrsg.), *Digital Libraries: Interdisciplinary concepts, challenges and opportunities. Proceedings of the Third International Conference on the Conceptions of the Library and Information Science, Dubrovnik, Croatia* (S. 97–108). Lokve.

Richardson, M. R., Brill, E. D., Ragno, R. J., & Rounthwaite, R. L. (2010). *Using popularity data for ranking* (Patent Nr. US007783632B2). https://patentimages.storage.googleapis.com/98/7f/87/8782ef18546fb8/US7783632.pdf

Rieh, S. Y. (2002). Judgment of information quality and cognitive authority in the Web. *Journal of the American Society for Information Science and Technology*, *53*(2), 145–161. https://doi.org/10.1002/asi.10017

Rieh, S. Y. (2009). Credibility and Cognitive Authority of Information. In *Encyclopedia of Library and Information Sciences, Third Edition* (Third ed., S. 1337–1344). CRC Press. https://doi.org/10.1081/E-ELIS3-120044103

Rieh, S. Y., & Belkin, N. J. (1998). Understanding judgment of information quality and cognitive authority in the WWW. *Proceedings of the 61st ASIS Annual Meeting*, 279–289.

Rieh, S. Y., & Danielson, D. R. (2007). Credibility: A multidisciplinary framework. *Annual Review of Information Science and Technology*, *41*(1), 307–364. https://doi.org/10.1002/aris.2007.1440410114

Roitero, K., Maddalena, E., Demartini, G., & Mizzaro, S. (2018). On Fine-Grained Relevance Scales. *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval – SIGIR '18*, 675–684. https://doi.org/10.1145/3209978.3210052

Ruthven, I. (2008). Interactive information retrieval. *Annual Review of Information Science and Technology*, *42*(1), 43–91. https://doi.org/10.1002/aris.2008.1440420109

Ruthven, I. (2014). Relevance behaviour in TREC. *Journal of Documentation*, *70*(6), 1098–1117. https://doi.org/10.1108/JD-02-2014-0031

Saracevic, T. (1975). Relevance: A review of and a framework for the thinking on the notion in information science. *Journal of the American Society for Information Science*, *26*(6), 321–343. https://doi.org/10.1002/asi.4630260604

Saracevic, T. (2007a). Relevance: A review of the literature and a framework for thinking on the notion in information science. Part II: Nature and manifestations of relevance. *Journal of the American Society for Information Science and Technology*, *58*(13), 1915–1933. https://doi.org/10.1002/asi.20682

Saracevic, T. (2007b). Relevance: A review of the literature and a framework for thinking on the notion in information science. Part III: Behavior and effects of relevance. *Journal of the American Society for Information Science and Technology*, *58*(13), 2126–2144. https://doi.org/10.1002/asi.20681

Saracevic, T. (2016a). Relevance: In search of a theoretical foundation. In D. Sonnenwald (Hrsg.), *Theory development in the information sciences* (S. 141–163). University of Texas Press.

Saracevic, T. (2016b). The Notion of relevance in information science: Everybody knows what relevance is. But, what is it really? In G. Marchionini (Hrsg.), *Synthesis Lectures on Information Concepts, Retrieval, and Services* (Bd. 8, Nummer 3). Morgan & Claypool. https://doi.org/10.2200/S00723ED1V01Y201607ICR050

Saracevic, T. (1996). Relevance reconsidered. In P. Ingwersen & N. O. Pors (Hrsg.), *Information science: Integration in perspectives. Proceedings of the Second Conference on Conceptions of Library and Information Science (CoLIS 2)* (Nummer CoLIS 2, S. 201–218). Royal School of Librarianship.

Saracevic, T. (2012). Research on relevance in information science: A historical perspective. In T. Carbo & T. Bellardo Hahn (Hrsg.), *Proceedings of the American Society for Information Science and Technology (ASIS&T) 2012 Pre-conference on the History of ASIS&T and Information Science and Technology* (S. 49–60). Information Today.

Saracevic, T. (2015). Why is relevance still the basic notion in information science? In F. Pehar, C. Schlögl, & C. Wolff (Hrsg.), *Re:inventing Information Science in the Networked Society. Proceedings of the 14th International Symposium on Information Science (ISI 2015)* (S. 26–35). Hülsbusch.

Sauro, J., & Lewis, J. R. (2012). Standardized Usability Questionnaires. In *Quantifying the User Experience* (S. 185–240). Jeff Sauro and James R. Lewis. https://doi.org/10.1016/b978-0-12-384968-7.00008-4

Savolainen, R. (2016). Information seeking and searching strategies as plans and patterns of action. *Journal of Documentation*, *72*(6), 1154–1180. https://doi.org/10.1108/JD-03-2016-0033

Savolainen, R., & Kari, J. (2006). User-defined relevance criteria in web searching. *Journal of Documentation*, *62*(6), 685–707. https://doi.org/10.1108/00220410610714921

Schamber, L. (1994). Relevance and information behavior. *Annual Review of Information Science and Technology (ARIST)*, *29*, 3–48.

Schamber, L. (1991). User's criteria for evaluation in a multimedia environment. *Proceedings of the 54th ASIS Annual Meeting*, 126–133.

Schamber, L., & Bateman, J. (1999). Relevance criteria uses and importance: Progress in development of a measurement scale. In L. Woods (Hrsg.), *Proceedings of the 62nd ASIS Annual Meeting* (Bd. 36, S. 381–389). Information Today.

Schamber, L., Eisenberg, M. B., & Nilan, M. S. (1990). A re-examination of relevance: toward a dynamic, situational definition. *Information Processing and Management*, *26*(6), 755–776. https://doi.org/10.1016/0306-4573(90)90050-C

Scholer, F., Kelly, D., Wu, W.-C., Lee, H. S., & Webber, W. (2013). The effect of threshold priming and need for cognition on relevance calibration and assessment. *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval – SIGIR '13*, 623–632. https://doi.org/10.1145/2484028.2484090

Schultheiß, S., Sünkler, S., & Lewandowski, D. (2018). We still trust in Google, but less than 10 years ago: An eye-tracking study. *Information Research*, *23*(3), paper 799. http://www.informationr.net/ir/23-3/paper799.html

Sedlmeier, P., & Renkewitz, F. (2018). *Forschungsmethoden und Statistik für Psychologen und Sozialwissenschaftler* (3. Aufl.). Pearson Deutschland.

Sheth, J. N., Newman, B. I., & Gross, B. L. (1991). Why we buy what we buy: A theory of consumption values. *Journal of Business Research*, *22*(2), 159–170. https://doi.org/10.1016/0148-2963(91)90050-8

Shokouhi, M., White, R., & Yilmaz, E. (2015). Anchoring and Adjustment in Relevance Estimation. *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval – SIGIR '15*, *3*, 963–966. https://doi.org/10.1145/2766462.2767841

Sims, D. B. (2002). *The effect of personality type on the use of relevance criteria for purposes of selecting information sources.* [University of North Texas]. http://digital.library.unt.edu/ark:/67531/metadc3313/m2/1/high_res_d/dissertation.pdf

Soergel, D. (1994). Indexing and retrieval performance: The logical evidence. *Journal of the American Society for Information Science*, *45*(8), 589–599. https://doi.org/10.1002/(SICI)1097-4571(199409)45:8<589::AID-ASI14>3.0.CO;2-E

Sperber, D., & Wilson, D. (1995). *Relevance: Communication and cognition* (2. ed.). Blackwell.

Surowiecki, J. (2005). *Die Weisheit der Vielen: Warum Gruppen klüger sind als Einzelne und wie wir das kollektive Wissen für unser wirtschaftliches, soziales und politisches Handeln nutzen können* (1. Aufl.). Bertelsmann.

Svenson, O. (1979). Process descriptions of decision making. *Organizational Behavior and Human Performance*, *23*(1), 86–112. https://doi.org/10.1016/0030-5073(79)90048-5

System Development Corporation. (1967). *Experimental studies of relevance judgments. Third progress report*. System Development Corporation.

Tague-Sutcliffe, J. (1992). The pragmatics of information retrieval experimentation, revisited. *Information Processing & Management*, *28*(4), 467–490. https://doi.org/10.1016/0306-4573(92)90005-K

Tang, R., Shaw, W. M., & Vevea, J. L. (1999). Towards the identification of the optimal number of relevance categories. *Journal of the American Society for Information Science*, *50*(3), 254–264. https://doi.org/10.1002/(SICI)1097-4571(1999)50:3<254::AID-ASI8>3.3.CO;2-P

Tang, R., & Solomon, P. (2001). Use of relevance criteria across stages of document evaluation: On the complementarity of experimental and naturalistic studies. *Journal of the American Society for Information Science and Technology*, *52*(8), 676–685. https://doi.org/10.1002/asi.1116

Taraborelli, D. (2008). How the web is changing the way we trust. In K. Waelbers, A. Briggle, & P. Brey (Hrsg.), *Proceedings of the 2008 conference on Current Issues in Computing and Philosophy* (S. 194–204). IOS Press. https://doi.org/10.5555/1566234.1566257

Taylor, A. R., Cool, C., Belkin, N. J., & Amadio, W. J. (2007). Relationships between categories of relevance criteria and stage in task completion. *Information Processing & Management*, *43*(4), 1071–1084. https://doi.org/10.1016/j.ipm.2006.09.008

Taylor, R. S. (1968). Question-Negociation and Information Seeking in Libraries. *College & Research Libraries*, *29*(3), 178–194.

Taylor, R. S. (1986). *Value-added processes in information systems*. Ablex Publishing Corporation.

Toepoel, V., & Funke, F. (2018). Sliders, visual analogue scales, or buttons: Influence of formats and scales in mobile and desktop surveys. *Mathematical Population Studies*, *25*(2), 112–122. https://doi.org/10.1080/08898480.2018.1439245

Tombros, A., Ruthven, I., & Jose, J. M. (2005). How users assess Web pages for information seeking. *Journal of the American Society for Information Science and Technology*, *56*(4), 327–344. https://doi.org/10.1002/asi.20106

Tombros, A., Ruthven, I., & Jose, J. M. (2003). Searchers' criteria For assessing web pages. *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval – SIGIR '03*, 385. https://doi.org/10.1145/860500.860513

Turpin, A., Mizzaro, S., & Maddalena, E. (2015). The Benefits of Magnitude Estimation Relevance Assessments for Information Retrieval Evaluation. *Proceedings of the 38th*

International ACM SIGIR Conference on Research and Development in Information Retrieval – SIGIR '15, 565–574. https://doi.org/10.1145/2766462.2767760

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, *185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Twait, M. (2005). Undergraduate Students' Source Selection Criteria: A Qualitative Study. *The Journal of Academic Librarianship*, *31*(6), 567–573. https://doi.org/10.1016/j.acalib.2005.08.008

Vakkari, P., & Hakala, N. (2000). Changes in relevance criteria and problem stages in task performance. *Journal of Documentation*, *56*(5), 540–562. https://doi.org/10.1108/EUM0000000007127

Vickery, B. C. (1959a). Subject analysis for information retrieval. *Proceedings of the International Conference on Scientific Information*, 855–866.

Vickery, B. C. (1959b). The structure of information retrieval systems. *Proceedings of the International Conference on Scientific Information*, 1275–1290.

Walraven, A., Brand-Gruwel, S., & Boshuizen, H. P. A. (2009). How students evaluate information and sources when searching the World Wide Web for information. *Computers and Education*, *52*(1), 234–246. https://doi.org/10.1016/j.compedu.2008.08.003

Wang, P. (1994). *A cognitive model of document selection of real users of information retrieval systems*. University of Maryland; College of Library and Information Science.

Wang, P. (2011). Information behavior and seeking. In I. Ruthven & D. Kelly (Hrsg.), *Interactive information seeking, behaviour and retrieval* (S. 15–41). Facet.

Wang, P. (2010). Contextualizing user relevance criteria: A meta-ethnographic approach to user-centered relevance studies. *Proceedings of the third symposium on Information interaction in context – IIiX '10*, 293–297. https://doi.org/10.1145/1840784.1840828

Wang, P., & Soergel, D. (1998). A cognitive model of document use during a research project. Study I. Document selection. *Journal of the American Society for Information Science*, *49*(2), 115–133. https://doi.org/10.1002/(SICI)1097-4571(1998)49:2<115::AID-ASI3>3.0.CO;2-1

Wathen, C. N., & Burkell, J. (2002). Believe it or not: Factors influencing credibility on the Web. *Journal of the American Society for Information Science and Technology*, *53*(2), 134–144. https://doi.org/10.1002/asi.10016

Watson, C. (2014). An exploratory study of secondary students' judgments of the relevance and reliability of information. *Journal of the Association for Information Science and Technology*, *65*(7), 1385–1408. https://doi.org/10.1002/asi.23067

Werner, K. (2019). *Benutzererwartungen: Eine interaktive Information Retrieval Studie zur Wahrnehmung von Suchergebnissen* [Universität Hildesheim]. https://doi.org/10.18442/016

White, H. D. (2009). Relevance in theory. In *Encyclopedia of Library and Information Sciences, Third Edition* (3rd Aufl., S. 4498–4511). CRC Press. https://doi.org/10.1081/E-ELIS3-120043266

White, H. D. (2019). Patrick Wilson. In *ISKO Encyclopedia of Knowledge Organization* (S. 1–50). https://www.isko.org/cyclo/wilson

Wildemuth, B., Freund, L., & Toms, E. G. (2014). Untangling search task complexity and difficulty in the context of interactive information retrieval studies. *Journal of Documentation*, *70*(6), 1118–1140. https://doi.org/10.1108/JD-03-2014-0056

Wilson, D., & Sperber, D. (2004). Relevance Theory. In L. R. Horn & G. Ward (Hrsg.), *The handbook of pragmatics* (S. 607–632). Blackwell.

Wilson, P. (1973). Situational relevance. *Information Storage and Retrieval*, *9*(8), 457–471. https://doi.org/10.1016/0020-0271(73)90096-X

Wilson, P. (1983). *Second-hand knowledge: An inquiry into cognitive authority*. Greenwood Press.

Xie, H. (Iris). (2008). *Interactive Information Retrieval in Digital Environments*. IGI Global.

Xie, I., & Benoit, E. (2013). Search result list evaluation versus document evaluation: similarities and differences. *Journal of Documentation*, *69*(1), 49–80. https://doi.org/10.1108/00220411311295324

Xie, I., Benoit, E., & Zhang, H. (2010). How do users evaluate individual documents? An analysis of dimensions of evaluation activities. *Information Research*, *15*(4), 1–21. https://doi.org/colis723

Xu, Y., & Chen, Z. (2006). Relevance judgment: What do information users consider beyond topicality? *Journal of the American Society for Information Science and Technology*, *57*(7), 961–973. https://doi.org/10.1002/asi.20361

Xu, Y., & Wang, D. (2008). Order effect in relevance judgment. *Journal of the American Society for Information Science and Technology*, *59*(8), 1264–1275. https://doi.org/10.1002/asi.20826

Xu, Y., & Yin, H. (2008). Novelty and topicality in interactive information retrieval. *Journal of the American Society for Information Science and Technology*, *59*(2), 201–215. https://doi.org/10.1002/asi.20709

Zhang, Y. (2014). Beyond quality and accessibility: Source selection in consumer health information searching. *Journal of the Association for Information Science and Technology*, *65*(5), 911–927. https://doi.org/10.1002/asi.23023